DOCUMENT RESUME

ED 181 050 TM 009 952

TITLE	Proceedings of the Invitational Conference on Testing Problems (2nd, New York, New York, October 29, 1949).
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	29 Oct 49
NOTE	86p.
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	Achievement Tests; Aptitude Tests; Conference Reports; *Cultural Factors; *Factor Analysis; Information Dissemination; *Information Needs; Intelligence Tests; Personality Tests; Psychological Studies; Psychological Testing; Research Design; *Testing Problems; *Test Results; Test Validity
IDENTIFIERS	*Testing Industry; *Test Reporting

ABSTRACT

The conference panels were organized around three topics: (1) influences of cultural background on test performance; (2) uses and limitations of factor analysis in psychological research; and (3) information which should be provided by test publishers and testing agencies on the validity and use of their tests. Panelists for the first session included Anne Anastasi, Ernest A. Haggard, William Stephenson, and William W. Turnbull. Panelists for the second session were George K. Bennett, H.J. Eysenck, and Paul Horst. Panelists and their topics for the third session were: Aptitude and Intelligence Tests, Herbert Conrad; Achievement Tests, Paul L. Dressel; and Personality Tests, Laurance F. Shaffer. Brief discussions followed each of the three panels. (MH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED181050

TM009 952

PROCEEDINGS

1949 INVITATIONAL CONFERENCE ON TESTING PROBLEMS

EDUCATIONAL TESTING SERVICE

PRINCETON, NEW JERSEY

LOS ANGELES, CALIFORNIA

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Educational Testing Service

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

EDUCATIONAL TESTING SERVICE

BOARD OF TRUSTEES

Herold C. Hunt, *Chairman*

Raymond B. Allen — Henry H. Hill

Joseph W. Barker — Katharine E. McBride, *ex officio*

Oliver C. Carmichael, *ex officio* — Thomas R. McConnell

Charles W. Cole — Lester W. Nelson

James B. Conant — Edward S. Noyes

George F. Zook, *ex officio*

OFFICERS

Henry Chauncey, *President*

Richard H. Sullivan, *Vice President and Treasurer*

William W. Turnbull, *Vice President*

Jack K. Rimalover, *Secretary*

Catherine G. Sharp, *Assistant Secretary*

Robert F. Kolkebeck, *Assistant Treasurer*

COPYRIGHT, 1950, EDUCATIONAL TESTING SERVICE,
20 NASSAU STREET, PRINCETON, N.J.
PRINTED IN THE UNITED STATES OF AMERICA

INVITATIONAL CONFERENCE ON TESTING PROBLEMS

October 29, 1949

OSCAR K. BUROS, *CHAIRMAN*

- Influence of cultural background on test performance.
- Uses and limitations of factor analysis in psychological research.
- Information which should be provided by test publishers and testing agencies on the validity and use of their tests.

EDUCATIONAL TESTING SERVICE

PRINCETON, NEW JERSEY ✦ LOS ANGELES, CALIFORNIA

FOREWORD

THE 1949 Invitational Conference on Testing Problems was enthusiastically received by those who were present. Since the excellent papers that were presented deserve a wider audience, they are being published in full, and along with them, the discussion from the floor that followed the formal presentations.

To Oscar K. Buros, who selected the topics, invited the participants, and conducted the meeting, goes full credit for the success of the conference. I would like to take this opportunity to express our grateful appreciation to him and to the speakers.

HENRY CHAUNCEY, *President*
Educational Testing Service

PREFACE

The 1949 Invitational Conference on Testing Problems, sponsored by the Educational Testing Service, was held at the Roosevelt Hotel in New York City on October 29, 1949. This conference was attended by more than two hundred educators, psychologists, and personnel workers interested in measurement and evaluation techniques.

In preparing the program, an attempt was made to select topics somewhat controversial in nature. Such topics appeared especially appropriate, since it has always been customary at the Invitational Conferences to allot considerable time for questions and criticisms from the audience. The topics were selected only after consultation with persons representing other viewpoints in testing; the final responsibility, however, for the selection of topics and speakers was my own.

The following three topics were selected for the conference program:

(1) Influence of Cultural Background on Test Performance
(2) Uses and Limitations of Factor Analysis in Psychological Research
(3) Information Which Should Be Provided by Test Publishers and Testing Agencies on the Validity and Use of Their Tests

It was felt that these topics represented a sufficiently wide range to permit all conference participants to find at least a part of the program of interest and value.

Speakers were selected so as to represent a variety of viewpoints. We were especially fortunate that two distinguished British psychologists, H. J. Eysenck and William Stephenson, were in this country at the time of the conference and agreed to present papers. An effort was made to select speakers who had not been on the conference programs in recent years. In retrospect, I think that I should have invited a representative of a test publisher to present a paper on "Information Which Should Be Provided by Test Publishers and Testing Agencies on the Validity and Use of Their Tests." This omission on my part is especially interesting since the conference was sponsored by the nation's largest test-construction and test-publishing organization. It speaks well for the Educational Testing Service that it made no attempt to influence me one way or the other in the selection of the topic for Panel III and in the selection of speakers.

I shall not attempt to summarize or assess the individual papers. In my opinion, all of the papers were of exceptionally high quality. This publication of the papers will permit others to evaluate the material for themselves.

I wish to express my gratitude to the speakers, to the discussants, to the numerous persons attending the conference, and to the Educational Testing Service for its sponsorship and efficient handling of the conference. I hope that the Educational Testing Service will continue to give us many more "Invitational Conferences on Testing Problems."

OSCAR K. BUROS, *Chairman*
1949 Conference

CONTENTS

PAGE

FOREWORD by Mr. Chauncey 5

PREFACE by Mr. Buros 7

PANEL I "*Influence of cultural background on test performance.*"

Anne Anastasi, *Fordham University* 13
Ernest A. Haggard, *University of Chicago* 18
William Stephenson, *Visiting Professor of Psychology, University of Chicago* 23
William W. Turnbull, *Educational Testing Service* 29

Discussion 35

PANEL II "*Uses and limitations of factor analysis in psychological research.*"

George K. Bennett, *Psychological Corporation* 41
H. J. Eysenck, *Institute of Psychiatry, University of London* 45
Paul Horst, *Educational Testing Service* 50

Discussion 57

PANEL III "*Information which should be provided by test publishers and testing agencies on the validity and use of their tests.*"

Aptitude and Intelligence Tests
Herbert Conrad, *U.S. Office of Education* 63

Achievement Tests
Paul L. Dressel, *Michigan State College* 69

Personality Tests
Laurance F. Shaffer, *Teachers College, Columbia University* 75

Discussion 79

Appendix 91

PANEL I

Influence of Cultural Background on Test Performance

Some Implications of Cultural Factors for Test Construction

ANNE ANASTASI

ANY discussion of the influence of cultural background on test performance involves at least two distinct questions. First, to what extent is test performance determined by cultural factors? Secondly, what shall we do about it?

In considering the first question, it is important to remember at the outset that culture is not synonymous with environment. Although this distinction should be obvious, some writers apparently forget it when drawing conclusions about heredity and environment. For example, environmental factors may produce structural deficiencies which in turn lead to certain types of feeblemindedness. Recent research on such conditions as Mongolism, microcephaly, hydrocephaly, and intracranial birth lesions has yielded a growing body of evidence for the role of prenatal environmental factors in the development of these conditions. Yet these types of mental deficiency would certainly not be classified as cultural in their etiology. Nor are they remediable in the individual case by education or by the manipulation of other cultural factors. Of course, the environmental factors leading to the development of these structural deficiencies may themselves be culturally influenced in the long run. Some day, we may know enough about them to control them through maternal nutrition, prenatal medical care, and the like. But such factors would represent an *indirect* cultural influence on behavior, mediated by structural deficiencies. Moreover, any such improvement in cultural conditions could have only a long-range effect and would not help the individual in whom the structural deficiency is already present.

Cultural factors do, however, affect the individual's behavior in many direct ways. Psychologists are coming more and more to recognize that the individual's attitudes, emotional responses, interests, and goals—as well as what he is able to accomplish in practically any area—cannot be discussed independently of his cultural frame of reference. Nor are such cultural influences limited to the more complex forms of behavior. There is a mass of evidence, both in the field observations of anthropologists and in the more controlled studies of psychologists, to indicate that "cultural differentials" are also present in motor and in discriminative or perceptual responses.

Now, every psychological test is a sample of behavior. As such, psychological tests will—and should—reflect any factors which influence behavior. It

9

is obvious that every psychological test is constructed within a specific cultural framework. Most tests are validated against practical criteria which are dictated by the particular culture. School achievement and vocational success are two familiar examples of such criteria. A few tests designed to serve a wider variety of purposes and possibly to be used in basic research are, in effect, validated against other tests. Thus when we report that a given test correlates highly with the number factor, we are actually saying that the test is a valid predictor of the behavior which is common to a group of tests. If we had no number tests in the battery, we could not have found a number factor. The type of tests which are included in such a battery—however comprehensive the battery may be—reflects in part the cultural framework in which the experimenter was reared. It is obvious that no battery samples *all* possible varieties of behavior. And as long as a selection has occurred, cultural factors are admitted into the picture.

In the construction of certain tests, special consideration has been given to cultural group differences in the selection of test items. The practices followed with regard to items showing significant group differences may be illustrated, first, with reference to *sex differences*. Insofar as the two sexes represent sub-cultures with distinct mores in our society, sex differences in item performance may be regarded as cultural differentials. The Stanford-Binet (1) is probably one of the clearest examples of a test in which sex differences were deliberately eliminated from total scores. This was accomplished in part by dropping items which yielded a significant sex difference in per cent passing. It is interesting to note, however, that it did not prove feasible to discard all such items, but that a number of remaining items which significantly favored one sex were balanced by items favoring the other sex. The opposite procedure was followed in the construction of the Terman-Miles Interest-Attitude Analysis (2), as well as in other similar personality tests designed to yield an M-F Index. In these cases, it was just those items with large and significant sex differences in frequency of response which were retained.

Another type of group difference which has been considered in the selection of test items is illustrated by the so-called *culture-free tests*, such as the International Group Mental Test (3), the Leiter International Performance Scale (4), and R. B. Cattell's Culture-Free Intelligence Test (5). In these tests, a systematic attempt is made to include only content which is universally familiar in all cultures. In actual practice, of course, such tests fall considerably short of this goal. Moreover, the term "culture-common" tests, would probably be more accurate than "culture-free," since at best, performance on such items is free from cultural *differences*, but not from cultural *influences*.

As a last example, let us consider *socio-economic level* as a basis for the evaluation of test items. One of the objectives of the extensive research proj-

ect conducted by Haggard, Davis, and Havighurst (6) is to eliminate from intelligence tests those items which differentiate significantly between children of high and low socio-economic status. On the other side of the picture, we find the work of Harrison Gough (7) in the construction of the Social Status Scale of the Minnesota Multiphasic Personality Inventory. In this scale, only those items were retained which showed significant differences in frequency of response between individuals in two contrasted social groups.

It is apparent that different investigators have treated the problem of cultural differences in test scores in opposite ways. An obvious answer is that the procedure depends upon the purpose of the test. But such an answer may evade the real issue. Perhaps it is the purpose of the tests which should be more carefully examined. There seems to be some practical justification for constructing a test out of items which show the maximum group differentiation. With such a test, we can determine more clearly the degree to which an individual is behaviorally identified with a particular group. It is difficult to see, however, under what conditions we should want to study individual differences in just those items in which socio-economic or other cultural group differences are lacking. What will the resulting test be a measure of? Criteria are themselves correlated with socio-economic and other cultural conditions. The validity of a test for such criteria would probably be lowered by eliminating the "cultural differentials." If cultural factors are important determiners of behavior, why eliminate their influence from tests designed to sample and predict such behavior?

To be sure, a test *may* be invalidated by the presence of uncontrolled cultural factors. But this would occur only when the given cultural factor affects the test without affecting the criterion. It is a question of the *breadth* of the influence affecting the test score. For example, the inclusion of questions dealing with a fairy tale which is familiar to children in one cultural group and not in another would probably lower the validity of the test for most criteria. On the other hand, if one social group does more poorly on certain items because of poor facility in the use of English, the inclusion of these items would probably *not* reduce the validity of the test. In this case, the same factor which lowered the test score would also handicap the individual in his educational and vocational progress, as well as in many other aspects of daily living. In like manner, slow work habits, emotional instability, poor motivation, lack of interest in abstract matters, and many other conditions which may affect test scores are also likely to influence a relatively broad area of criterion behavior.

Whether or not an item is retained in a test should depend ultimately upon its correlation with a *criterion*. Tests cannot be constructed in a vacuum. They must be designed to meet specific needs. These needs should be defined in advance and should determine the

choice of criterion. This would seem to be self-evident, but it is sometimes forgotten in the course of discussions about tests. Some statements made regarding tests imply a belief that tests are designed to measure a spooky, mysterious "thing" which resides in the individual and which has been designated by such terms as "Intelligence," "Ability Level," or "Innate Potentiality." The assumption seems to be that such "intelligence" has been merely overlaid with a concealing cloak of culture. All we would thus need to do would be to strip off the cloak and the person's "true" ability would stand revealed. My only reaction to such a viewpoint is to say that, if we are going to function within the domain of science, we must have operational definitions of tests. The only way I know of obtaining such operational definitions is in terms of the criteria against which the test was validated. This is true whether a so-called practical criterion is employed or whether the criterion itself is defined in terms of other tests, as in factorial validity. Any procedure, such as the discarding of certain items, which raises the correlation of the test with the criterion, enables us to give a more precise operational definition of the test. But we cannot discard items merely on the basis of some principle which has been laid down *a priori*, such as the rule that items showing significant group differences must be eliminated. If this procedure should lower the validity coefficient of the test, it could have neither practical nor theoretical justification.

It is also pertinent to inquire what would happen if we were to carry such a procedure to its logical conclusion. If we start eliminating items which differentiate subgroups of the population, where shall we stop? We could with equal justification proceed to rule out items showing socio-economic differences, sex differences, differences among ethnic minority groups, and educational differences. Any items in which college graduates excel elementary school graduates could, for example, be discarded on this basis. Nor should we retain items which differentiate among broader groups, such as national cultures, or between preliterate and more advanced cultures. If we do all this, I should like to ask only two questions in conclusion. First, what will be left? Secondly, in terms of any criterion we may wish to predict, what will be the validity of this minute residue?

REFERENCES

(1) McNemar, Q. *The Revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin Company, 1942. Pp. 185. (cf. Ch. V)

(2) Terman, L. M., and Miles, C. C. *Sex and Personality: Studies in Masculinity and Femininity.* N.Y.: McGraw-Hill, 1936. Pp. 600.

(3) Dodd, S. C. *International Group Mental Tests.* Princeton Univ. Ph.D. Disser., 1926. Pp. 101.

(4) Leiter, R. G. *The Leiter International Performance Scale.* Santa Barbara, Calif.: Santa Barbara State College Press, 1940. Pp. 94.

(5) Cattell, R. B. A Culture-Free Intelligence Test: I. *J. Educ. Psychol.*, 1940, Vol. 31. Pp. 161-179.
Cattell, R. B., Feingold, S. N., and Sarason, S. B. A Culture-Free Intelligence Test: II. Evaluation of Cul-

tural Influence on Test Performance. *J. Educ. Psychol.*, 1940, Vol. 32. Pp. 81-100.

(6) Haggard, E. A., Davis, A., and Havighurst, R. J. Some Factors which Influence Performance of Children on Intelligence Tests. *Amer. Psychol.*, 1948, Vol. 3. Pp. 265.

Davis, A., and Havighurst, R. J. The Measurement of Mental Systems (Can Intelligence Be Measured?). *Scient. Monthly*, 1948, Vol. 66. Pp. 301-316.

(7) Gough, H. G. A New Dimension of Status: I. Development of a Personality Scale. *Amer. Sociol. Rev.*, 1948, Vol. 13. Pp. 401-409.

Influence of Culture Background on Test Performance

ERNEST A. HAGGARD

In considering the topic of the influence of culture background on test performance, I will limit myself this morning to some of the factors which influence the mental test performance of children in our society. In this connection, we are all convinced that how well a child does on an intelligence test is a function, in part at least, of a complex of genetic factors. But there is a good deal of disagreement on the extent to which they are reflected by scores on our present mental tests. In a strictly factual sense, no one can say. About all that we can be sure of is that there is no conclusive and definite relation between inheritance and performance on mental tests in terms of such variables as, for example, socio-economic class.

This has been, and still is, a focal point of many ardent and heated controversies. But because of the great number of genes that underlie the inheritance of the higher mental processes, and because of the large number of generations of discrete stratified samples that would be necessary to demonstrate the specific inheritance of such processes, and because of the relatively rapid movement of individuals up and down the socio-economic ladder in America, it seems unlikely that any of us will see an empirical solution of this problem. Nor may we find a solution in the past. In the last few centuries in western European culture, there has been no stable stratification of intelligence in terms of socio-economic class, because of social upheavals like revolutions and wars. Probably the most stable group is that of royalty, the topmost level, but it is questionable whether this group has distinguished itself for intellectual achievement. The theoretical and mathematical solutions arrived at by such geneticists as Haldane, Hogben and Huxley indicate no demonstrable difference in the inheritance of mental abilities in different socio-economic groups.

In addition to genetic factors, individuals also inherit, in a sense, a physical and social environment. The effects of these may be tested. For example, from studies with animal and human subjects, we already know something of the effects of serious nutritional deficiencies on the development of neural and other bodily tissues, and their impairment of later adaptive behavior. Such early deficiencies and weaknesses may also lay the groundwork for various debilities in later life. The recipients of such handicaps are

characteristically found in the lower socio-economic groups. If, then, a lower-class child were perfectly "normal" at birth—he still may, because of various factors which impede normal development, actually be subnormal by the time he is of testable age. But perhaps more important is the emotional deadening, the development of mental callouses, the disinterest in life, and loss of willingness to respond to it, that often accompanies severe deprivation. Little careful work has been done in this area, except perhaps for some of the studies of institutionalized and rejected children, and some of the foster home and twin studies. In any argument referring back to inheritance, however, such factors must be considered.

Everyone is aware of the influence of the social or cultural environment on test performance. No one would think of giving an intelligence test standardized on American children to a child in Bali, or France, or South Africa—and expect the results to mean very much. No one would give such a test to a child on the other side of the ocean, but few have accepted the fact that results of such testing might be invalid if it is given to a child on the other side of the tracks. This is a basic point, and the source of much misunderstanding and faulty "information." Again, in terms of any genetic argument, one must consider the fact that children from privileged (or middle-class) homes receive a range of experiences and acquire a range of motivations which prepare them much more adequately for favorable performance on our present type of intelligence tests than is the case with lower-class children.

The thesis here is that we cannot assume the various sub-cultures in America to be comparable, simply because of a common geographical boundary. Nor is it enough to say "We consider all that when we evaluate the IQ of a lower-class or ethnic child." Why not? Because all too often the educational opportunities—from the early grades on—are determined by how well a child does on our present standardized tests, regardless of whether we intend them to be. And those who do poorly at first are often given inferior educational opportunities, so that a vicious circle is set up, and a great deal of potential ability lost to our society.

Furthermore, test-constructors cannot, by themselves, design tests which are equally fair to all children in our society. One reason for this is that they, themselves, are middle-class individuals, buttressed by middle-class experiences, ways of thinking, and language forms. It is not surprising, then, that they construct tests which are saturated with middle-class vocabulary and language forms, and that the experiences and knowledges tested are those with which most middle-class children are relatively familiar, and that they often standardize their tests in terms of such middle-class values as academic achievement. The problem then is whether these customary procedures are really appropriate for testing the mental abilities of children reared in a lower-class culture. From

15

all we know of the wide range of differences between socio-economic groups in America, the answer is No.

The way to approach such a problem is to find out what the lower-class and ethnic cultures in our society are like. The best techniques for finding out about other cultures is by first making anthropological and sociological field studies. A sizable number of such investigations have been made of various social and ethnic groups in America. Allison Davis, and the people working with him at the University of Chicago, have made use of these findings in their attempt to construct culturally-common intelligence tests for American children. It is manifestly impossible to have a culturally-free test, since symbol systems must be used, but we feel that it is possible to base a test on a range of experiences which is sufficiently common to all children in our society.

But before describing some of our procedures and findings, I would first like to indicate one or two possible effects stemming from an incomplete awareness of the many differences between, say, lower-class and middle-class groups. Generally, it seems that it has too often been assumed that a test would yield an adequate "measure" of intelligence because of the sheer elegance of the statistical computations involved. Such procedures are necessary, but are not a sufficient justification for an intelligence test. One must also examine carefully what each item measures, as well as the proportion and distribution of subjects passing an item, or a battery of items, and the correlation with some criterion, at a given age level. On examination, a great majority of items found in the present intelligence tests is biased in favor of the middle-class child, and against the lower-class child. This is due largely to the particular types of information sampled, the vocabulary and verbal forms used, as well as the artificial and academic nature of many of the item types themselves. An additional difficulty with many present tests is that they are standardized on a somewhat biased sample in terms of the total population. There are more middle-class children in school—especially at the older ages, it is much easier to motivate them to take the tests, and in general they are more cooperative and pleasant to work with. But if the testing conditions, the items, the test as a whole, or the sample on which they are standardized are biased in favor of middle-class children, then it follows that lower-class children would necessarily obtain lower "IQ" scores—even though they really were as gifted.

The research program of Davis and others at Chicago is directed toward the development of individual and group intelligence tests which are maximally fair to all social-class and ethnic groups in our society, that is to say, a culturally-fair intelligence test. There have been roughly four steps in this research program.

The first step involved extensive and intensive anthropological field studies, in which middle-class and lower-class ethnic and white groups were studied. Some graduate students lived in the

homes of lower-class families for several months. Children from all groups were studied in their homes, schoolrooms, and neighborhoods. These children were interviewed on how and on what they spent their time, their range of experiences, how they used words, what words meant to them, how they solved test problems, and what things were important to them.

The second step involved an examination of the relative performance of some 5,000 children on each of the 460 items from ten frequently used intelligence tests. From these data attempts were made to find out why large discrepancies existed in certain items and item types. It was frequently due to artifacts of middle-class verbal habits, or differences in background experiences, or differences in motivation to do well on the tests. It furthermore appeared that if the tests had been equally fair for children from extreme social-class groups in terms of familiarity and motivational factors, the wide performance differences would have been greatly reduced or wiped out.

The third step involved an experimental demonstration that such factors as differential familiarity, motivation, and the removal of middle-class artifacts did significantly reduce the difference in the performance of lower-class and middle-class children on the same mental test problems. In a pilot study, it was found that the differential performance of the two extreme social-class groups could be modified almost at will—either in terms of increasing or decreasing this difference. In a later and more intensive study, 636 children took part in a five-day experiment investigating the effects of the following variables on mental test performance: social-class, practice, motivation, the form of the test, and its manner of presentation. The complete set of findings is too numerous and complex to report here, but there are a few which are especially worth mentioning. Between an initial test and a retest (with three practice periods on similar item types intervening), (a) the lower-class children showed as great an over-all gain in performance as did the middle-class children; that is, they learned or profited from their experiences as much as the middle-class children, (b) the lower-class children, when highly motivated on a retest of standard-type intelligence test items did significantly better than the lower-class children not thus motivated, (c) many items were revised to remove the middle-class bias, and (d) the children from both social-class groups profited more from the experimental conditions with which they were more familiar in terms of previous experience, training, etc. But even though many traditional item types can be reworked to be less discriminating against lower-class children, without violating the essential nature or difficulty of the item, it was felt that in general they were too academic and artificial, and that a new approach should be taken, with the development of items which are not only fair in terms of the background experiences of all groups, but equally motivating as well.

The fourth step, which is being carried on at present involves the construction and standardization of a new battery of individual and group intelligence tests for children of ages six to nine inclusive. Since this step is only partially completed, I am not able to describe the specific tests. However, according to a statement by Davis and Hess, the item types used in the new tests include: the understanding of physical principles, the classification of objects into categories selected by the child, memory processes, the drawing of inferences from given relationships, critical processes and the ability to verify solutions, general inductive and deductive reasoning, and a number of others. The items themselves involve problems and problem types which are about equally common to all socio-economic groups —but problems which are not taught either in the home or the school. Consequently, they have to be solved almost entirely by the child's reasoning and creative ability, as it may be developed by his general experience.

The group tests are not completely analyzed as yet, but findings from the individual tests show a lack of difference in the average performance of children from extreme socio-economic groups on these non-academic problems. Yet at the same time, adequate age distributions were obtained, as were correlations between this test and school achievement scores for each socio-economic group. The magnitudes of these correlations are comparable to those reported in the literature for the present intelligence tests.

In conclusion, I would like to return briefly to the question of the relation between the inheritance of mental ability and socio-economic differences in intelligence test performance. In this connection, two points seem relevant: first, that the intelligence test scores are, in themselves, irrelevant data so far as any proof or disproof of genetic theories are concerned, and second, on test problems common to children from many socio-economic and ethnic groups, but problems which are not dependent on specific school experience, we did not find the customary distribution of mental test scores in terms of socio-economic level. Therefore, the burden of proof for demonstrating that the upper socio-economic groups inherit a complex of gene characteristics which are tied to superior mental ability, and lower socio-economic groups inherit inferior genetic structures, rests on the shoulders of those who interpret such differences in mental test scores as being due to differential inheritance.

Influence of Cultural Background on Test Performance

WILLIAM STEPHENSON

BY CULTURAL BACKGROUND, however defined, one refers to historically-rooted matters. A white-collared Eton scholar and a back-street Brooklyn boy appear to be distinguishable, and yet also indistinguishable, in terms of their cultural backgrounds. Their sentiments, habits, attitudes, and affections may well be very different. Yet both speak English, and both live with a common heritage of law, religion, customs, and much else besides. They are educated differently, yet the ideals of Ancient Greek and Renaissance Educators penetrate into the schools of both. In comparison with a Hindu, however, or a sedate Chinese boy for whom Taoism was a background until Communism burst in upon him, Europeans seem culturally different. And still more diverse must be the culture in and against which the native African child lives, or an Eskimo.

It has been difficult, however, if not impossible, to formulate concrete and operational postulates about such cultural agglomerations. It is permissible, perhaps, to distinguish between (a) educational influences, (b) socio-dynamic situations, and (c) the vague, historically determined culture patterns which, when evaluated, we grace with the name of heritage, and with which we are here to be concerned.

One might have thought that "culture" psychologists, whose very problem was to interpret these latter historical trends along the lines laid down by their mentor, Dilthey (1), would have provided something for us to bite upon, scientifically, by now. True, they produced an Oswald Spengler, with his notable *Decline of the West*, but I know of no testable hypotheses that reach into Spengler's Mayan, Babylonian, Graeco-Roman, or any other "civilizations." Yet interesting matters are at issue. We know that the ancient Athenians, after the Persian Wars, created the European mind out of a mere handful of human beings and a few square miles of territory; and our own Elizabethan Golden Age, after a hundred years of war, was born almost within sight of London. Moreover, whereas the Greeks and the Elizabethans called for the richest development of a man's *general* ability, the current trend is rather to foster the trickling specialties and presumed aptitudes of our young. So that perhaps culture determines very largely *what* abilities we shall value and develop, rather than anything else at issue. There are strong

19

suspicions that this is the case, as sociologists such as Mannheim, for example (2), have already suggested.

Up to very recently, however, testing, such as we are to consider, has hinged upon a *null* hypothesis. This is to the effect that cultural differences have little or no effect upon some really important dimensions of human personality. It is implied that there may be only a few such dimensions, perhaps only one, or two. We find the hypothesis almost unexpectedly, wherever we turn; at bottom it represents a belief that there must be general laws of personality which transcend cultures—and by laws I mean theories, or synthetic propositions as Kaufmann and many modern philosophers would call them (3), which serve as models or growing points for hypotheses that can be put to experimental test. This null hypothesis lies behind the search for so-called "culture-free" intelligence tests: and indeed it would surely be imposing, if not important, if it could be shown that individuals drawn from widely different contemporary cultures, such as our English, African, American and Chinese boys, are alike in certain important essentials.

If this null hypothesis has finally to be rejected, we may still wonder whether there are on record any clear instances where important personality features have been shown to have for the main part a cultural determination. The possibilities of any essential *interactional* standpoint, however, can perhaps be discounted; for it scarcely seems reasonable to suppose that there can be much interaction between an ordinary person and his cultural milieu, such that each influences the other and everything is relative to everything else. For the individual is surely a puny speck against his cultural background. Exceptions to this, of course, are the great men and women of culture, a Plato, Aristotle, Buddha, or the like.

What have test performances, then, to say about these various matters? We should put aside, I think, any consideration of studies relating to heredity, or to the influence of socio-economic levels upon test performance, since these, except as controls, are scarcely pertinent to the questions at issue concerning culture.

Consider the null hypothesis first. One may begin by wondering whether a Kinsey Report for widely diverse national and cultural groups would read very differently in essentials from the American. Or, if we distinguish between *thinking* and *intelligence*, as Bartlett would have us do, interesting findings such as those of Carmichael (4) come to light. Using a verbal-projective test consisting of unfinished newspaper editorials on controversial topics, Carmichael showed that Cambridge graduates and English working-class men and women, all alike, intelligent and unintelligent, argued illogically, rationalized quite naively, projected and generally played havoc with anything that resembles the orderly procedures of an intelligence test. Would not the same apply the world over? Or consider another example. Thematic Apperception tests may well mirror the immediate behavioral stresses, strains,

and preoccupations of different individuals, and to this extent very obvious social and perhaps culturally-determined differences may be brought to light. But if Sam has trouble with his wives, past and present, and Alexandrovic with his party affiliations, and Nagawooli with his goat—who is interested in these matters as such? One may interpret the results, of course, perhaps psychoanalytically, and so point to basic affinities of a dynamic kind underlying all these preoccupations. It may be shown in this way, for example, that children in slum areas appear to have far severer superegos than children from better-to-do homes (5). But the psychoanalyst might well demur about such an apparent result, pointing out that only superficial indications of psychoanalytical dynamics are tapped by such tests, and that greater penetration might, rather, show everyone, of all cultures, alike in essentials: thus, the psychoanalyst, too, becomes involved in a null hypothesis for his fundamental postulates.

Along systematic lines, however, the best example I can offer is from work in the Spearman School. This began with a distinction (made on theoretical grounds which were rooted in late English Associationism) between *noetic* and *anoetic* processes. The former was represented formally by Spearman's g-factor, and the latter by all manner of specifics and group factors within the cognitive field of study.[1] Line next showed that "visual perception" in children paralleled their mental growth, that is, their mental age (against which, of course the Binet tests had been validated originally). Stephenson (6), and Brown and Stephenson (7) followed by indicating that tests of this same visual perceptual material could be regarded as "pure" tests of Spearman g-factor, with these noetic implications. Finally, Fortes (8), who turned from the London group to become an anthropologist, found that African natives performed this kind of perceptual test quite as satisfactorily as whites. Fortes, however, was careful to do what others rarely achieve in test construction: he randomized the varieties and styles of perceptual material by selecting it from every known culture, past and present.[2]

Now I make no claim that this se-

[1] It is one of the sad consequences of a purely inductive approach to factor work that Burt, Thurstone, and most text-books, refer to the Spearman Theory of Two Factors without reference to the experiential matters and psychological theory that the factor theorems merely echoed, or paralleled as models. Thus, Spearman merely wished to deny the proposition that group factors could be found in the *noetic* field; he knew full well that they abounded in the *anoetic*.

[2] Stuart Dodd (9) attempted something of this kind for pictures of common objects and situations, for his so-called international test of intelligence. But the materials and problems were rooted in *anoetic* processes, and the test showed greater rather than less differences between racial groups. Similarly the styles of the fundaments used in the Penrose-Raven matrices (10), and in Cattell's (11) "culture-free" test, or Penrose's new perceptual test, are severely European and geometrical in form, and to this extent would be suspect wherever the null hypothesis *wasn't* supported. They would be suspect for other reasons, too, but I must leave this to one side for the present.

quence of events and its outcome was other than tentative: it lacked the resources for test construction and standardization that America now affords, or that the Educational Testing Service so elegantly devotes to its tests. But its theoretical implications were clear, and obviously it was orientated towards this null hypothesis. Moreover, I propose not to enter into the appraisal of such results as we have available about "culture-free" or any other tests involving us in this null hypothesis: there is some evidence, such as that of Fortes, supporting the hypothesis for perceptual data, and much purporting to reject it. In the latter cases, however, so little has been done, usually, to randomize materials, or to take account of other controllable factors, that the evidence is at least dubious. I can only suspect that Fortes and the Spearman School were at least on the right lines to handle mainly visual perceptual material for some kind of crucial test of the null hypothesis.

But now let us consider the other proposition, that culture has a decided, even a decisive, effect on human personality. For most of us this may seem completely obvious. It is surely easy enough to bring different national attitudes to light, as Cantril (12) is perhaps doing. Here I would like to be pardoned for using my own experimental observations, since I believe that they are methodologically more at the heart of what is involved.

I begin with the knowledge that it is the *type* psychologists, the Sprangers and the Jungs, and sociologists such as Mannheim (2) and Fromm (13) in recent years, who stress the influence of cultural background on present personality. But it appears that no self-respecting psychometrist, except myself, believes any more in *types*, except as cuts across a normal distribution, made for convenience—much as we cut up the I.Q. scale into moron, feeble-mind, normal, and genius. Even so, I would ask you to re-open the whole matter of types, or at least to keep an open mind about it for the next few years, for I believe the psychometrists have been barking up quite the wrong tree. Matters look very different if one approaches types from a Q-technique standpoint (14).

It is a simple matter, for example, to show that more men in the United States are likely to be of a type X, that we might call "extrovert," than of a type Y, that we might call "introvert." The opposite is the case for women. But the main types can be demonstrated for any small number of persons, for example for any ten of you in this room, without operational reference to any other persons in or out of the room. Indeed we can say something about the matter for only *one* person if need be: thus, given a "population" of 200 traits chosen at random from a Jungian universe of such traits (I have 2,000 traits in such a universe), I might invite the *one* person (a) to appraise himself with the traits, (b) then, having done this, to give an account of what he believes an ideal introvert to be, and (c) finally to give an account of what he believes an ideal extrovert to be. The correlations

between (a), (b), and (c) for N=200 traits, will indicate whether our one person (if he is sophisticated like ourselves or college students) is of introverted or extroverted type.[3]

But for the moment we need only examine the implications of such Q-technique findings, and its approach, for our preoccupation with culture. Suppose that, in terms of Q-technique, types are now demonstrable (as indeed they are). In the case of Introversion-Extroversion such types were rooted, for Jung (15), almost wholly in cultural background. Jung traced the matter back into pre-Christian history; into the disputes and castigations of a Tertullian and an Origen of some eighteen centuries ago; into Schiller's idealization, many centuries later, of the "Grecian heaven"; into the massive folklore and poetry of a *Faust*, a *Parsifal*, or a *Zarathustra*; and so down into the very tough mindedness of James's *Pragmatism*.

Now it may stretch one's credulity, if not one's imagination, to accept the proposition that these same roots find their way into the personality of our one person whose correlations have just been referred to. Yet clearly he operated with my little test, and it is not really difficult to see that his evaluations of the traits may very well stem just precisely into or from such historically persistent strands.

At the outset he was asked merely to give a description of his own personality in terms of the 200 innocuous-looking traits. He had no idea that I was going to ask him, subsequently, to describe an ideal or typical introvert and an extrovert. Nor did the traits suggest that anything of the kind was likely to be involved. Clearly some kind of ostensible learning has mediated, and the culture psychologist was perhaps quite correct to trace this not only into current culture (plus learning in an ostensible manner), but also to seek its roots in cultural history.

The psychometrist, however, has not sought to represent such types but to measure isolated, perhaps a-historical or immediate, functions or factors, such as introversion-extroversion or the like—much as one measures an electric current. At best the result has been not one function or factor, but several, to judge for example from Guilford's studies. One doubts, however, whether anyone feels happy about these factors, for they really do not explain anything, they are incapable of consequent operational tests, and indeed different forms of analysis can provide rather different apparent factors.

The situation is very different if one seeks to *represent* types as such statistically. For one can then operate with the types, that is, subject them to experimental tests, even for only one person at a time.

One can see the fashioning of such

[3] Thus, for the following quite typical data, the person is very likely to be introverted in type (or thinks he is):

	Self (a)	Ideal I (b)	Ideal E (c)
(a)	—	+50	−55
(b)		—	−90
(c)			—

types, interestingly enough, in current American culture. Eric Fromm (13) for example, in his *Man for Himself*, offers a description of the supposed "market" type of personality, which he ascribes to Americans who apparently want to sell everything, including their own personalities. In terms of Q-technique I have recently reduced Fromm's notions to some kind of orderly operational testing, and can readily demonstrate, and thus verify, his "market" characterization of Americans. This, apparently, is fashioned by your culture.

But what we prove is that such-and-such men are *alike* in type. It is quite another matter to test them for any underlying functions in terms of individual differences. By the very postulates one uses, in the latter case, one throws away any possibility of achieving concrete types as such.

In conclusion, then, cultural influences can be brought into full view in the *typification* of human beings, as Spengler, Jung, and others down to Fromm have seen. I state it as a testable postulate that any systematic quantification in terms of individual differences (which we are unfortunately wont to regard, almost as a myth, as the exclusive concern of our testing procedures) cannot represent such typification, and certainly is in no way needed for its achievement.

As I see the issues, therefore, in the very broadest manner I am prepared to examine the null hypothesis that cultural background is neutral, or can be randomized, with respect to some of our major psychological preoccupations. These are functions such as noesis, libido, and the like. As an offshoot, it is perhaps as well to remember that society also determines what abilities will be valued, and what discounted. But by the same token it is now easy to demonstrate that man's personality *types* are fashioned very probably in terms of the culture in which he lives.

REFERENCES

(1) Hodges, H. A. *Wilhelm Dilthey: an Introduction.* N.Y., 1944.
(2) Mannheim, K. *Man and Society.* N.Y., 1949.
(3) Kaufmann, F. *Methodology of the Social Sciences.* N.Y., 1944.
(4) Carmichael. *British J. of Psychology (Gen. Sec.)*, 1943.
(5) Jackson, L. *Ph.D. Thesis, University of Oxford*, 1948.
(6) Stephenson, W. *J. Educ. Psychol.*, 1931, Vol. 22.
(7) Brown, W., and Stephenson, W. Chapter VII of Brown and Thomson, *Essentials of Mental Measurement.* N.Y., 1940.
(8) Fortes, M. *Ph.D. Thesis, University of London*, 19[illegible].
(9) Dodd, S. C. *International Group Mental Tests.* Ph.D. Thesis, Princeton University, 1926.
(10) Penrose and Raven. *British J. Med. Psych.*, 1943.
(11) Cattell, R. B. A Culture-Free Intelligence Test. *J. Educ. Psychol.*, 1940.
(12) Cantril, H. UNESCO Studies, 1949.
(13) Fromm, E. *Man for Himself.* N.Y., 1949.
(14) Stephenson, W. (see bibliography in Wolfle, D. *Factor Analysis to 1940.* Chicago U.P.)
(15) Jung, C. G. *Psychological Types.* N.Y., 1935.

Influence of Cultural Background on Predictive Test Scores

WILLIAM W. TURNBULL

For convenience in attacking the broad question before this panel I should like to limit my discussion to tests used for the purpose of prediction. In imposing this limitation, however, I feel that I am not greatly restricting the field of inquiry, since in the final analysis most test scores derive their utility from their predictive significance.

If we consider tests whose use is frankly predictive, two questions of interest are: first, when people of different cultural backgrounds take the same test, how do their scores compare? And second, are differences in level of test performance of different cultural groups associated with similar differences in the subsequent behavior those scores were supposed to predict?

In an approach to the first of these questions, Henry Chauncey and I carried out some years ago a study (as yet unpublished) to discover the manner in which students from different geographical areas and different sizes of communities differed in their performance on types of questions commonly used in tests of scholastic aptitude.

The test used was the first Army-Navy College Qualifying Test. This examination included four sections, consisting of verbal, scientific, reading, and mathematical material, respectively. The verbal section included questions relating to word meaning, word usage, and the like, in the form of opposites, analogies, and definitions in completion form. The second or "scientific" section, was composed of questions of the so-called common-sense science type. The technical information needed to answer them was not great, and for the most part intelligent scientific interest and alert observation would prove as valuable as scientific training. The third section of the test consisted of paragraphs of rather general nature, each followed by questions on its content, while the mathematical section (the last section of the test) was designed to test numerical reasoning, presupposing a background of arithmetic, elementary algebra, and rudimentary geometry.

The test was given in 1943 to over 300,000 students all over the country, as a screening device for the college training programs of the Army and Navy. All of the people tested were male, were 17-21 years of age, and had reached or passed the senior year of secondary school.

From the mass of answer sheets, eight subgroups were segregated, first

by taking all answer sheets from the four regions of New York State, Alabama and Georgia combined, Iowa and Nebraska combined, and California; and then by separating within each region the answer sheets of students in large and in small communities. A large community was defined as one whose population was 150,000 or more and a small community as a non-suburban community below 5,000 in population. For a convenience these groups were called urban and rural respectively. (I will readily agree that these terms are not rigorous, since not all students attending school in a community of fewer than 5,000 souls come from farm homes, although a substantial proportion of them do.) Finally, from each of the eight groups (two sizes of community within four geographical areas) a random sample of 500 answer sheets was drawn.

Please note particularly that the samples were far from random or representative samples of the total student population of the age range 17-21 in the four regions. They represent merely the extremes on a scale of population size, within groups that had voluntarily taken the qualifying tests; and there is no basis for ascribing representativeness to the samples.

The main results of this study

ARMY–NAVY COLLEGE QUALIFYING TEST
RURAL–URBAN DIFFERENCES BY REGIONS

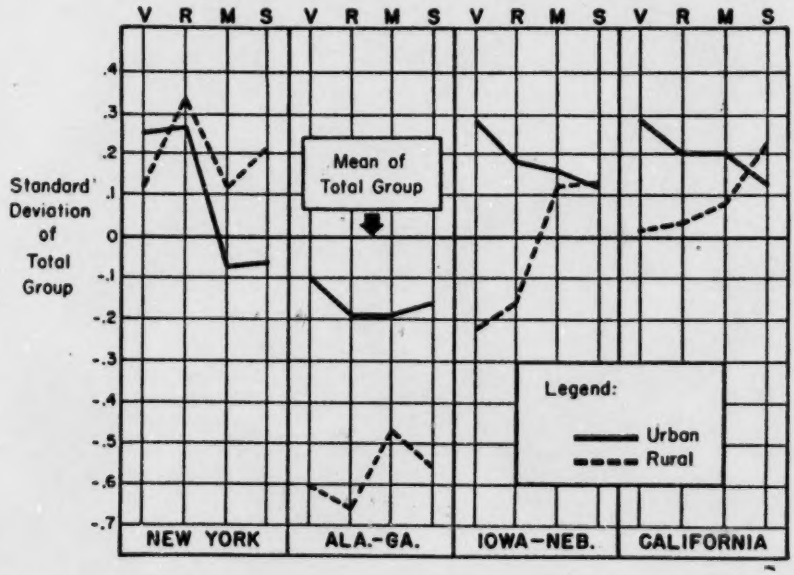

FIG. I

26

have been put in graphic form, and I believe there are sufficient copies here for each person present. Looking first at the sheet headed "Army-Navy College Qualifying Test—Rural-Urban Differences by Regions," (Fig. 1) we see first the very conspicuous depression of all values for Alabama-Georgia, particularly in the rural areas (represented by the dotted line). Since the scale here is expressed in tenths of a standard deviation for the total group, it is evident that the rural Alabama-Georgia candidates scored about three-fourths of a sigma below their rural New York cousins. For the urban groups the regional differences are less striking, but are still present. Next notice that the solid lines tend to slope *downward* to the right, while the dotted lines tend to slope *upward* to the right. This is seen clearly from the chart where the composite rural-urban comparison is made (See Fig. 2), with all four regions averaged together. Evidently the students from large communities were much more facile verbally than those from small communities, whereas in mathematical ability their superiority was slight, and in terms of ability to answer common sense science questions the two groups were equal.

An analysis of variance showed that the differences in total test performance according to geographical region

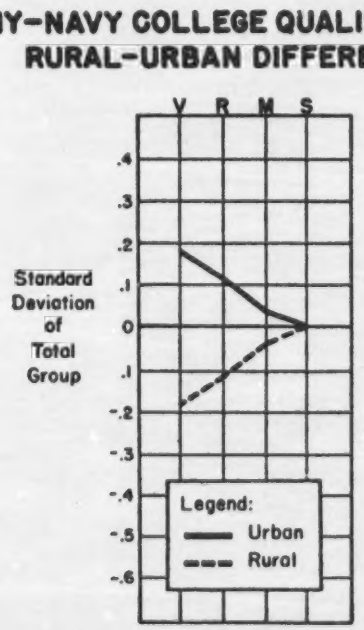

FIG. II

were statistically significant, as were the differences in performance according to size of community. There were significant differences between states in the relationship of the ability of people from large communities to that of people from small communities. And finally, rural-urban differences varied significantly according to the kind of test material used, as illustrated in the second graph.

As a further step in this investigation, separate item analyses were completed for the eight subgroups on ten items from each test section, in an attempt to discover whether the lower scores of the rural group resulted from generally poorer performance on the items within a given test section or from failure on particular items. I shall not take time to report in detail on our findings, but an analysis of variance showed that the item difficulty differences between the groups varied significantly from one item to another. That is, the order of item difficulty was not the same for boys from small communities as for boys from large cities. In the case of the verbal section of the test the variance of item difficulties by community sizes was considerably larger than would have been required for significance at the 1% level of confidence, and for the other three sections was significant at better than the 5% level of confidence. Similarly, the differences in performance of geographically distinct groups varied significantly from item to item.

The fact that the differences between groups depend on the individual test questions considered, rather than merely on the type of test material, is of crucial importance for the argument as to the cause of the differences. For if the differences were common to all items of one type or one factor, we might argue that the different groups had inherited different patterns of abilities and that these patterns reflected themselves directly in the test section scores. But one would scarcely argue for differential inheritance of ability to solve individual questions within such a homogeneous factor as verbal, where the dependence of rural-urban differences on the particular test question asked was most clearly indicated. The conclusion must be that, whether or not there were inherited mental differences between our groups from large and small communities, and from state to state, environmental differences must have caused certain of the differences in test performance: specifically, the intergroup differences in order of item difficulty within a single test section.

If one grants that some test questions are relatively harder than others for people in a specified cultural group, the next question is: what shall we do about it? Should we build tests that minimize the intercultural difference in scores, or that maximize that difference, or shall we trust to chance to bring us out somewhere in the middle?

It is my contention that on a predictive test any score difference between groups whose backgrounds differ should be judged not good or bad, not right or wrong, but useful or not useful, valid or invalid for the prediction of future behavior. We must

specify the criterion we wish to predict, and then justify intergroup equality or inequality of test scores on the basis of its effect on prediction.

Relatively little attention has been given to the question of the effect on prediction of score differences between cultural groups. The results of a few investigations are available, and they show in general that the rather haphazard mixing of items favorable to various subcultures has so far resulted in tests that differentiate usefully among cultural groups, if one's purpose is to predict the criteria used in these studies.

In a study shortly to be published by Frederiksen and Schrader a comparison was made between predicted achievement and actual achievement of veteran and non-veteran students in their first year of college. In four institutions a prediction of the freshman grades was made from scores on an aptitude test, using the ACE Psychological Examination in three instances and the College Board Scholastic Aptitude Test in the fourth. Within each group, veteran and non-veteran, a division was then made on the basis of background variables, for the purpose of discovering whether or not they were associated with a tendency to accomplish more in college than the test scores predicted. If the aptitude tests were assigning improperly low scores to students of lower socio-economic status one would expect such students to over-achieve (in relation to predictive test score) on the criterion variable, whatever it might be. Such was not the case in the four institutions studied. Background data were available on income of family head, formal education of father, and size of community. No clear trends emerged to show that the student's position relative to these variables was related to his tendency to over-achieve, whether veteran or non-veteran students were considered.

In the study of the College Qualifying Test whose results I have reported no criterion data were obtained. Such data were, however, gathered in an unpublished study by Conrad and Robbins, who used the same qualifying test to predict achievement after two semesters of the V-12 program. They then attempted to account for the errors in prediction on the basis of educational handicap in high school using as the measure of educational handicap the average teacher's salary in the school system from which the individual came. The hypothesis tested was that teacher's salary should correlate negatively with over-achievement: the lower the salary, the greater the excess of achievement over prediction. Out of seventeen colleges studied, negative correlations were found in six, a zero correlation in one, and positive correlations in ten, showing that whatever educational handicap was reflected in the aptitude scores was reflected to at least as great degree in first year college achievement.

These findings suggest that intergroup differences on scholastic aptitude tests, when the grouping is based on factors usually associated with cultural or educational handicap, are valid for

the prediction of college freshman grades. Admittedly this is a limited criterion, but the nature of validation demands that we investigate our criteria one by one.

The findings based on freshman grades were corroborated in a further 8-college study reported by Conrad and Robbins at the 1947 meeting of the American Educational Research Association. In that study the authors found that errors in prediction of *fifth-term* college work from aptitude test scores were not related either to average teacher's salary or to size of community from which the student came: that is, whatever handicap the factors of teacher's salary or community size may reflect manifested itself as strongly in achievement through the fifth college term as it did in the aptitude scores obtained before entrance to college.

I wish I could report findings of similar studies aimed at longer-range criteria of greater social significance. Unfortunately, however, I know of no existing data that will help us answer the question of the validity, for such criteria, of the intergroup differences under consideration.

To summarize, the study of intergroup differences on the Army-Navy College Qualifying Test, reported earlier in this paper, illustrates the magnitude of the score differences obtained when one administers a typical scholastic aptitude test to groups of high-school graduates in various geographical regions and from communities of different sizes. The analysis of differences on individual test questions points to the casual influence of cultural differences in producing score differences. Other investigations have uncovered evidence that such score differences have some predictive utility: i.e., that when college grades through the fifth term are accepted as a criterion, the test scores reflect accurately the performance of the various subgroups. What we need in order to provide a more generally useful answer to the question of predictive utility are studies in which test scores are used to forecast long-term life success. Only studies of this kind can tell us how great should be the intergroup differences on predictive tests.

DISCUSSION

PARTICIPANTS:

OSCAR K. BUROS, ANNE ANASTASI, WILLIAM STEPHENSON, HAROLD GULLIKSEN, ERNEST HAGGARD, HUGH M. DAVISON, WILLIAM W. TURNBULL, DOUGLAS E. SCATES.

CHAIRMAN BUROS: First I will give the members of the panel an opportunity to raise any questions they have with the other members of the panel.

DR. ANASTASI: I actually agree with what has been said by most of the speakers. I would like to make three points in this connection: First of all, I think that studies of cultural differences in test performance are extremely important. The sort of study that Dr. Turnbull has just reported and that Drs. Haggard, Davis, and Havighurst have done on how cultural groups differ in test performance, is very important in helping us to understand what the existing differences are and to what extent cultural background affects performance.

I think, too, that studies of cultural similarities by such tests as Dr. Stephenson mentioned, studies that are using tests which are culturally neutral or culturally randomized and which enable us to focus our attention, therefore, on what these cultures have in common, are also important.

I believe, however, that such studies are quite apart from the problem of constructing tests. When we construct tests, I would agree thoroughly with Dr. Turnbull that the criterion is the only thing we can go by. As for the use of tests for purposes other than prediction, I still say that we must provide an operational definition of what we are testing. And unless we have an operational definition in terms of a criterion, I do not know what it would be.

DR. STEPHENSON: I always like the chance to say an extra word, if I may say just one.

I should hate to think that I leave you with the impression that technical matters are of no consequence. Clearly, if you are making tests for practical purposes, these are important. I would merely like to support Dr. Turnbull to that extent. I found results similar to those described by Turnbull for the British Army and Air Force. When I took V and G and K tests, the G test did not differentiate men and women, for instance, in the armed forces, on some eighty thousand samples, but the V test certainly did so.

The K test was so badly done by women that it was almost unbelievable.

I therefore know that these facts are there to look for, but I should still be wondering whether I shouldn't plan it out in some way, you see; why should I be looking for just V and R and S and M? There should be some main dimensions involved, and it is there that I think a little theory rather than a mere scramble would, perhaps, help.

Otherwise, I have nothing but admiration, as I said, for the elegance, and so forth, for which all the technical matters of test construction are accounted. I still have my problem that there are some very big issues, like the one of the Greek culture and the Golden Era; they are a phenomenon that has happened, and it would be so nice if we could find something that would alter things now so that we might have another sort of era; that would seem like a completely fantastic dream, though, I know.

CHAIRMAN BUROS: The meeting is open to questions from the floor.

DR. GULLIKSEN: Mr. Chairman, I was interested in the emphasis on validity from the speakers, and I want to ask Dr. Haggard a question. I thought I detected one sentence in his talk that dealt with the question of validity. To what extent have your studies dealt with not only the validity of your test for different socio-economic groups, but also the validity of the older tests which you are criticizing? And how do these validities of the two types of test compare for different economic groups, for predicting various criteria? Could you give a summary of such findings, please?

DR. HAGGARD: Yes, but the data that will enable me to answer your question are just coming in, and are being analyzed at present, so my remarks must be somewhat general.

As one measure of validity for each social class group, we used a reading achievement test, since it was the one test given by the school system to all the children in the study. Our individual intelligence test predicted performance on reading achievement as well as, or slightly better than, the Kuhlman-Anderson and the Primary Mental Abilities tests for each social class group.

The use of such criteria for determining validity, however, has not been our main interest. Rather, we had in mind an approach which may appear a little naive—namely, reliance on face validity. In other words, we selected items which met the criterion, "This test item should be passed by a smart boy and failed by a stupid boy, regardless of his socio-economic status." The decision for the inclusion or exclusion of an item was made by a group of qualified experts in such fields as education, psychology, and anthropology.

If, for each social class group, our tests can predict school achievement as well as present intelligence tests, it may be said that we don't need to worry. But in saying this, I mean to place worry in quotation marks, because we don't believe that such a measure is a very meaningful criterion

of validity, although it is about all we have if we must rely on such objective data. We don't like it because the usual criterion measure is heavily biased in favor of middle-class children. In spite of this, we found correlations around .50 between our tests and the measure of school achievement—which was about the same as that for the standarized tests with our subjects. A correlation of this size, however, leaves enough variance unaccounted for so that, even though our tests do not correlate with socio-economic status, they may still correlate as highly as standard tests with some such "valid" criterion. We believe that since social class bias is minimized in our tests, they actually provide a more nearly valid measure of intellectual ability.

CHAIRMAN BUROS: Are there any other questions?

DR. DAVISON: I should like to address these remarks to Dr. Turnbull. I think his graph is very interesting, but I wonder if it does not show that New York and California are strong school systems rather than the existence of a city-rural difference. Also, there is this possibility, that in New York you have proportionately more centralized school districts, so that you really do not have a typical rural school situation. Furthermore, the curriculum is clear in New York and California, maybe more so than in the other states, although I believe the southern states are picking up now in the matter of what they teach in the schools. However, these data have gone back into the past, because these people have grown since public school days and there may be a change in the coming youth.

There was a recent study done in Kentucky, in Caldwell County, I believe, and their picture looked something like this graph of the Alabama and Georgia group you have here. Maybe you can call it a matter of culture, but I think perhaps it is a matter of education more than culture.

DR. TURNBULL: I would have no quarrel with that. I think I feel there is no basis in the data for separating education from other cultural aspects. There is one additional variable that may be causing the differences found in these graphs; that is, the brighter students from the rural areas in Alabama and Georgia may not, for some reason, have volunteered to take the college qualifying test. Differential selection probably was operative. Whether or not it was in that direction, I am not prepared to say. But I should not like to leave the impression that these results are thought of as representing the educational systems of the four regions, since that distinction is blurred by the self-selection that took place before the tests were given.

DR. SCATES: May I say a word with respect to Dr. Anastasi's emphasis on the correlation with the criterion as a measure of validity. I cannot go as far as she does on that point. We must recognize that the criterion itself has some "bugs" in it; that the criterion is just as difficult to define and sometimes more so than the trait in which we are immediately interested. I know

it is a very simple, neat, and, in the abstract, a logical thing to say that we will set up tests which correlate well with a criterion, but usually this criterion is not the well defined thing we assume it is when we glibly refer to it.

The research process operates in both directions, not just one. The tests we set up frequently serve in helping to redefine the criterion by throwing light on its complex structure. They may prove also to have various forms of utility of their own and receive justification in part for this reason. In the case of intelligence, we are not trying to obtain and define a trait solely to predict any one particular thing; we are trying to define, and gradually refine our concept of, a trait because we are interested in a workable concept of intelligence itself. In the process of thus establishing a useful trait, we desire to know its many characteristics such as correlation with various things. Some of these things may be regarded more or less as criteria. But the trait being measured has rights of its own, and correlations with various criteria, while furnishing indexes of certain utilities, do not constitute the sole measure of the essential nature of the trait being measured.

CHAIRMAN BUROS: Do you wish to reply to that, Dr. Anastasi?

DR. ANASTASI: Yes, just briefly. I would like to say that I fully agree that the criterion has "bugs" in it, and I wanted to emphasize just that. When the criterion has "bugs" and we validate a test against that criterion, then the test has those "bugs," too, and we must not forget that the test has them. If we call it an intelligence test, that label will not eliminate the "bugs." If we define intelligence and then forget that definition in the process of constructing the test or validating it, we have not thereby eliminated the "bugs." That is just why I want to focus attention upon the criterion.

DR. SCATES: To accomplish this end we must give attention to the "bugs" in the criterion as well as those in the trait. We cannot properly place our emphasis solely on either one or the other.

DR. ANASTASI: I do not know what a "trait" means in such a case.

PANEL II

Uses and Limitations of Factor Analysis in Psychological Research

Uses and Limitations of Factor Analysis in Psychological Research

GEORGE K. BENNETT

ALTHOUGH the title of this morning's panel discussion is "The Uses and Limitations of Factor Analysis in Psychological Research," I think I should make it clear at the beginning that I am going to talk about only a limited portion of this topic. My concern will be primarily with the production of useful test batteries and the contribution that is made to them by factor analysis. However, I should like to mention briefly some general notions about factor analysis which seem to be pertinent to this particular application.

To many people one of the great appeals of factor analysis is its apparently solid foundation in mathematical theory. To a certain extent this is a valid belief. The problem of factor analysis is, from a geometric viewpoint, the problem of finding the minimum number of reference axes needed to describe a distribution of scores. Whereas the original references are the tests, the new references are the factors, and there are to be fewer factors than tests. This problem is clearly mathematical in nature, and to seek its solution is consistent with the scientific principle of parsimony. However, once the reference axes have been determined, the process ceases to be mathematical. The identification or naming of factors and the use of the factors or the results of factorial analysis in practical psychological work is no longer a mathematical problem. From here on we are concerned with such questions as the extent to which the resulting factors have been influenced by the composition of the initial battery of tests, the characteristics of the sample from which the data were obtained, the applicability of the results to other groups, the violations of psychological and mathematical theories in converting the results to practical and feasible testing procedures, and finally the rather simple question: Now that we have factorial results, what are we going to do with them?

Coming to the actual process of the construction of a battery of tests by means of factorial analysis, the steps are something like this: Since no factors can eventually be obtained which are not included among the variables initially studied, factorial analysis ordinarily begins with a large number of tests. It is desirable that these tests represent the largest possible variety of abilities, and furthermore, that each test be a reasonably pure one. "Purity," in this connection, refers to homogeneity of content and process. Inasmuch as relatively few pure tests have been in

general use, the factor analyst often finds it desirable, if not necessary, to construct new tests for this purpose. As we all know, test construction is a time-consuming and expensive process, particularly when one utilizes conventional methods of item analysis to construct a power test in which the items have high correlation with total score and are arranged in order of difficulty. Furthermore, power tests usually consume extended periods of time, and in a situation where as many as 60 tests are to be administered to each subject, the total time required can reach prohibitive lengths. Consequently, we find the major portion of original test batteries in these situations consisting of highly speeded tests of relatively simple functions in which the score depends largely on the number of attempts made per unit of time.

After the matrix of correlation coefficients has been obtained, the initial factor loadings are computed and, according to some factor analysts, the axes should be rotated so that the number of zero loadings is maximized so that the factors shall make sense. Although the initial factors by definition have no correlation with each other, the rotated axes often are not entirely independent; in other words, there is some sacrifice of independence for the sake of improved factor identification. If practical use is to be made of the factorial results, it is necessary that the test battery be abbreviated to manageable lengths. This usually means not over three or four hours of testing time, and the pressure from teachers and school guidance personnel makes even shorter times more advantageous. This means that the original battery of 60 tests must be reduced to a much smaller number, say twelve or fifteen as a maximum. This involves selecting those tests which, either singly or in combination, will yield the best estimate of each factor. If we have as many as six factors, this means that no more than two or three tests can be used to identify each, unless a particular test is weighted separately for different factors. Since the correlation of particular tests with the first factors extracted tends to be high, reasonably good identification of two or three factors will ordinarily result, but some factors often have rather low loadings in any test with subsequent poor estimation from any combination of a small number of tests. Although the factors are correlated only to a small extent, the individual tests are usually much more highly correlated, and score combinations from these tests equally so. This leads to the situation reported by Crawford and Burnham[1] among others in which the average correlation of Thurstone's PMA battery is reported as .36, whereas the Yale battery, not constructed on a factorial basis, yields an average intercorrelation of only .41. This would appear to be a very small gain in independence of score for the factorially constituted battery.

A far more important defect, which is not due to the factorial process per se, is the substitution of "factorial" validity for real or practical validity. So far as I know, the authors and

publishers of factorially constructed test batteries have been satisfied to report factorial validity and to imply that these are adequate substitutes for what they somewhat condescendingly refer to as "practical validity."[2] From the standpoint of usefulness to the counselor, factorial validity is a wholly inadequate substitute. The counselor is faced with the necessity for making a series of differential predictions in order to estimate the degree of success and satisfaction which his client may expect in each of the several courses of action that are feasible for him to undertake. In order to make such decisions, the counselor needs to know the extent to which his test scores are important to success in certain school courses and jobs. While the persons who have constructed factorial batteries have not been unaware of this need and have listed occupations with which, in their judgment, the various factorial scores may be expected to have positive relationships, it is my belief that this is the flimsiest sort of conjecture, inasmuch as no experimental data are brought forth in support of these contentions. If it is reasonable to expect that the authors and publishers of non-factorial test batteries should produce evidence of validity against realistic criteria, it appears also reasonable to require evidence of the actual validity of the factor scores resulting from factorially constructed batteries. This is particularly true in the case of less well defined factors since often these do not coincide even approximately with any traits for which some evidence of validity has previously been obtained. For example, what interpretation can one give to a factorial score on the basis of such a definition as this:

> "This second factor seems, then, to represent a difference between spatial ability, perhaps combined with some numerical ability and a certain amount of manual dexterity, and verbal ability, with possibly some memory involved. As a further speculation we suggest that it might be linked with some physiological or temperament factor which gave mental activities a singleness of direction and resistance to change very similar to the inertia of a moving body."[3]

This statement was made in reference to an unrotated factor, but equally vague statements, in perhaps fewer words, have occasionally been made with regard to rotated factors.

This brings us to the problem of what can be done to make factorial analysis more useful in terms of test battery construction. The first and most obvious step would be to undertake a series of realistic validational studies to determine the extent to which each of the factor scores is predictive of success in various scholastic courses and occupational categories. It may well be that singly and in combination factorial scores have definite advantages over the scores from a good battery of tests constructed according to more traditional principles. I very much doubt that this will be the case, but I am willing to admit the possibility. A much more realistic application of factorial analysis would include a number of criterion scores among the variables initially studied. If these criteria could represent a realistic sam-

pling of several quite different scholastic or occupational activities, the resulting knowledge could have very extensive significance for measurement and perhaps for educational philosophy. It is probable that educational situations ordinarily do not offer the opportunity for obtaining criteria of this sort for any adequate number of individuals. It might, however, be possible to set up an experimental school in which several quite different types of training could be offered within the span of one academic year to a large number of students. Objective and comprehensive proficiency tests would be required for each type of training so that reliable and meaningful criterion scores could be obtained. A factor analysis of tests and criteria for these students would then result in factors which would have meaning in specific situations. It might be found that some factor scores are suitable for predicting a large number of criteria or it may be found that criteria which are apparently very much alike require different factors for adequate prediction. Whether we use conventionally-made tests or factor scores or projective methods, we must still establish validity as the power to predict for specific groups and for specific criteria. This type of experiment might ultimately result in the extraction of factors which simplify the counselor's task and effect considerable economy in testing time. On the other hand, it might indicate there is no great value in factor scores. But until we try the experiment, we won't know the answer. In view of the fact that great sums of money are being spent on educational experiments, the cost of such an undertaking would not seem to be prohibitive.

Lacking the types of validational evidence that have been briefly suggested in the preceding paragraphs, it is my belief that factorial analysis has not demonstrated any unique values in terms of test battery construction, although it has given us some useful clues to mental organization and has, perhaps, provided some reinforcement of the notion long ago stressed by Kelley, Hull and others that, if several tests are to be used in combination, low intercorrelations are desirable.

REFERENCES

(1) Crawford, A. B., and Burnham, P. S. *Forecasting College Achievement.* New Haven: Yale University Press, 1946. Pp. 200.

(2) Guilford, J. P. *Manual of Instructions and Interpretations for the Guilford-Zimmerman Aptitude Survey.* Beverly Hills, Calif.: Sheridan Supply Co., 1947.

(3) Wolfle, Dael. Factor Analysis to 1940. *Psychometric Monog.*, 1940, Vol. 3. Pp. 12.

Uses and Limitations of Factor Analysis in Psychological Research

H. J. EYSENCK

FACTOR analysis has been much criticized by orthodox statisticians as well as by idiopathically-minded psychologists, although for different and frequently opposite reasons. These criticisms often stem from inadequate understanding—inadequate understanding of the assumptions involved and the statistical methods used on the part of the "idiopaths," and inadequate understanding of the purposes underlying its use on the part of the statisticians. As always, the uses and limitations of a mathematical method of analysis depend on the purposes which it is designed to serve. In the case of factor analysis, there appear to be two main purposes: 1) To discover taxonomic principles in a field in which so little is known that no reasonable hypotheses can be set up and tested, and 2) to test deductions made from taxonomic hypotheses in a field studied sufficiently to allow the setting up of promising theories. In both cases, it will be seen, the problem is one of taxonomy or classification; factors are conceived as principles of classification which allow us to order our field of study in a way determined by the properties of the material with which we are dealing, rather than in terms of subjective preference, intuition, or on the basis of common sense.

It will be clear that factor analysis is differentiated from all the orthodox procedures of statistics—determination of significance of differences, analysis of variance and covariance, discriminant function analysis, sequential analysis, and so forth—by the fact that where all the orthodox procedures test the null hypothesis as regards differences between certain groups which are known *a priori*, or in terms of previous experimental investigation, factor analysis attempts to answer the much more fundamental question: "What are the principles of classification which obtain in this particular field, and according to which experimental groups ought to be selected for the determination of significant differences?"

This differentiation links up with the fundamental problem in mental testing, namely that of validity. We must distinguish very clearly between two types of validity, which we may tentatively call lower-order validity and higher-order validity. The usual textbook definition of validity as "agreement with a criterion" refers to lower-order validity, and is essentially an engineering concept. If we select

a simple criterion, such as number of bolts soldered per hour, or number of accidents per month, we can easily determine the "validity" of a given test by correlating it with the criterion. But while such a determination may have a certain amount of practical usefulness in human engineering, its scientific value is almost precisely nil.

The difficulty of this conception of "validity" is brought out clearly when we apply it to truly psychological concepts such as "intelligence," or "extraversion," or "suggestibility." Here we have either no criterion at all, or a multiplicity of criteria which do not correlate very highly with each other. We must therefore look for a criterion to decide which of the many criteria to use, a procedure which gives rise to an infinite regress of looking for criteria to decide which of several criteria to use in deciding which is the correct criterion, and so forth.

Once this situation arises in which no clear-cut external criterion is available—and this is the case in connection with every genuinely psychological concept I know of—we must have recourse to some form of higher-order concept of validity. Such a concept can only derive from the adoption of the internal-consistency approach; in other words, as in every other science, the isolated fact acquires meaning only in relation to other facts, and interpretation, measurement, and conceptualization become possible by coordinating the isolated facts in a system capable of functional development through the use of the hypothetico-deductive method (1). It is not claimed that factor analysis is the only possible variant of this internal-consistency approach; it is merely claimed that at the present stage of development of mental testing procedures, no other method is available which will answer the taxonomic, classificatory questions which arise. Nor is it claimed that factor analysis is perfect in its present form; all of us who have used it on any large scale will agree that there are many aspects of it which require improvement or even drastic overhauling. But for the type of problem I have outlined, there simply does not appear to be an alternative, although that does not mean that we should not go on looking for one which is free of the admitted difficulties attending the factorial approach.

Two brief examples will illustrate the use of factorial methods in relation to the two main purposes I mentioned at the beginning. The first related to the discovery of taxonomic principles in a field in which so little is known that no reasonable hypothesis can be set up and tested. In our early work on factors determining aesthetic preferences we made an attempt to discover the reasons underlying preferences for different types of poetry (2). The literature threw no light on this problem, and consequently a factor-analytic design was set up. Some thirty poems, each relatively short, were ranked in order of preference by our subjects; these rankings were correlated and factor analyzed. Two factors emerged, without rotation, which could be interpreted very clearly on the basis of the poems most liked and

most disliked by the subjects having high positive or negative saturations respectively on these factors. The first factor divided those who like a simple rhyming scheme (abab), a regular, evenly accentuated rhythm, and a clearly defined ending to each line from those who like complex rhyming schemes, irregular, uneven rhythms, and lines that continue from one to the other without clear breaks. I do not want to waste time by discussing the second factor also; one factor will illustrate my point sufficiently. Starting from a position in which we have no guiding principles as to how we should classify our material, we emerge with a clear-cut hypothesis determined essentially by the internal organization of the preference judgments. This hypothesis allows of disproof and of functional development; we have tested it by predicting preference for poems not contained in our original sample, and we have developed it functionally by showing that similar principles of organization obtain in preferences for pictures, jokes, statues, and other aesthetic objects, and by showing that this simplicity-complexity factor is correlated with temperament (3). Many other examples could be given, but I think this one is sufficient to illustrate our point that factor analysis may give rise to classificatory hypotheses.

As an example to illustrate our second claim, namely that factor analysis can be used to test a classificatory hypothesis, I may perhaps quote our studies in suggestibility (4). The hypothesis was set up that eight well known tests of suggestibility measured one and the same underlying variable, which might be identified with this concept of "suggestibility." Intercorrelations were run between the eight tests, and a factor analysis performed. The analysis showed that two factors were needed to account for the observed correlations within the limits of the sampling error, and that the tests were grouped in the two-dimensional space of this two-factor pattern in such a way that four tests—the body sway test originated by Hull, the Chevreul Pendulum test, and two arm levitation tests—constituted one group, while tests of the Binet type—progressive lines, progressive weights, etc. —constituted the other group. These two groups were entirely uncorrelated, the angle of separation between the centroids passing through them being almost exactly 90°. The original hypothesis is conclusively disproved and the hypothesis suggested that we are dealing with two separate types of suggestibility which we called "primary" and "secondary," or "ideomotor" and "sensory" suggestibility. When this new hypothesis was tested, by using different populations, and additional tests, such as a measure for hypnotizability and a variety of tests of the sensory kind, the deductions made were confirmed in each instance (5). Here then we have an example of how factor analysis can be used to disprove a hypothesis, namely that of a general factor of suggestibility, how it can suggest instead another hypothesis, and how it can be used to test this new hypothesis.

A third possible use of factor analysis may lie in a field in which it has not hitherto been used to any significant extent, namely that of the description of social groups. It is customary to describe individuals and groups in terms of scores on psychometric tests; thus a group of democrats may be more "radical" than a group of republicans in terms of some measure of radicalism-conservatism. However, it is possible that differences between groups may be apparent more in the organization of component attitudes than in over-all scores. Two groups may not differ with respect to "radicalism" as measured, but they may show differences with regard to the pattern of intercorrelations between the component attitudes. Factor analysis appears to be the preferred method for disclosing and quantifying such differences in organization, and it has been used in this way in our studies into the organization of social attitudes as determined by political party, by age, sex, education, and by nationality.

If these are the uses of factor analysis, what are its limitations? One serious limitation lies in the lack of statistical criteria of significance for factor loadings, for variances, and for residuals. While we have approximations, and at least one method, namely that of Lawley, which permits of the application of such criteria, nevertheless the absence of practicable and accurate methods for estimating significance is a serious business. Another limitation is implied in the outline of the use of factor analysis given above—the evidence given by factorial methods is often suggestive rather than definitive, permissive rather than conclusive. However, factor analysis shares this limitation with almost all other scientific research methods.

While these limitations are admitted, others, also often suggested by critics, are not. The fact that factor analysts do not always agree, for instance, is no more a criticism of factor analysis, and does not set up any more necessary limitations, than does the fact that the respective schools started by Weierstrass and Kronecker in mathematics hold diametrically opposed views on the nature of such a fundamental concept as numbers, limit the usefulness of mathematics. The fact that factor analysis makes certain assumptions regarding linearity and the additive nature of its variables does not constitute a necessary limitation, as these assumptions can be tested, and as methods of factor analysis not dependent on them can be envisaged. The fact that factorial analyses often give results which are plainly absurd constitutes a limitation of factorial analysis only in the sense that this statistical method does not guarantee success, when inappropriately used and inexpertly handled, any more than does calculus or any other mathematical technique. Factor analysis requires just as insightful statement of the problem, just as careful design of the experiment, and just as skillful interpretation psychologically of the results, as does any other technique; if used in any

other way it will prove misleading and unhelpful. Its limitations, insofar as they are not merely of a temporary technical nature, are defined by its purposes—while useful and indeed essential for certain purposes, it throws no light on other types of problems, and does not attempt to displace other methods more adequate for their solution. In other words, like all scientific methods, its usefulness is not universal, but circumscribed, and only the mature judgment of the expert can decide whether in a given situation it is likely to give him the answer he wants.

REFERENCES

(1) Eysenck, H. J. Dimensions of Personality. N.Y.: Macmillan, 1949.
(2) ———. Some Factors in the Appreciation of Poetry, and Their Relation to Temperamental Qualities. *Char. and Person.*, 1940, Vol. 9. Pp. 160-167.
(3) ———. The Experimental Study of the "Good Gestalt"—A New Approach. *Psychol. Rev.*, 1942, Vol. 49. Pp. 344-364.
(4) ———. Suggestibility and Hysteria. *J. Neurol. Psychiat.*, 1943, Vol. 6. Pp. 22-31.
(5) ———, and Furneaux, W. D. Primary and Secondary Suggestibility: An Experimental and Statistical Study. *J. Exp. Psychol.*, 1945, Vol. 35. Pp. 485-503.

Uses and Limitations of Factor Analysis in Psychological Research

PAUL HORST

It seems to me that scientific investigations in any discipline must be concerned primarily with the variables which are thought to be fundamental to the science or discipline. If there is general agreement on what the fundamental variables of the science are, then it is difficult to see how factor analysis of any sort could be of value. However, if the investigators within a discipline cannot generally agree on what these variables are, then I believe factor analysis can play a useful role in providing an objective basis for agreement. Certainly in psychology there is a marked lack of agreement with reference to those variables important for describing, predicting, and controlling human behavior.

But in what sense may we regard any set of variables as basic for the science? In general, the factor problem arises whenever in a given discipline there exists, first, a large number of entities belonging to a specified class, and second, a large number of experientially distinct attributes on the basis of which the entities are differentiated from one another. In particular, the entities may be people and the attributes may be tests. Again, the entities may be corporate stocks and the attributes may be successive days for which the prices of each of the stocks are given. Or the entities might be various geographical regions and the attributes might be variables such as wind velocity and direction, relative humidity, barometric pressure, and other atmospheric variables.

Factor analysis assumes that there exists a relatively small number of attributes on the basis of which the entities may be differentiated from one another about as adequately as on the basis of the large number of experiential attributes. Factor analysis assumes that the numerical values of the experiential attributes can be expressed as functions of the primary variables. The traditional and current methods of analysis have assumed these functions to be linear but these assumptions are not necessary except for practical convenience.

Essentially, then, the purpose of factor analysis is to simplify in a very specific manner our description of observable phenomena. This simplification is achieved by reducing the number of attributes we require to differentiate one entity from another. Presumably, we may regard the primary attributes as a subgroup of the larger group of experiential attributes, if we include not only those which have already been

numerically evaluated, but also those which might conceivably be so evaluated in the future. It can be demonstrated, however, that if one relatively small group of attributes exists such that all the others may be expressed as functions of this group, then there will also exist an infinite number of such groups, even though they may not all be experimentally independent.

The question therefore arises as to which of the many possible small groups of attributes one should select as a basis for estimating the large number of experiential attributes. Again, we shall adopt the criterion of simplicity of description and define it in numerical terms. Let us assume that we have approximately 30 attributes in each subgroup, and that all of the 30 are needed to estimate all of, say, 10,000 experiential attributes, within the limits of accuracy we impose upon ourselves. Assume, however, that for one of these groups of 30 we need an average of only 15 of the attributes to predict with acceptable accuracy each of the experiential variables. For another group we need an average of 20 of the primary variables to estimate each of the 10,000 experiential variables. We find, however, finally that for a particular group of fundamental variables we need only an average of 7 of the 30 to estimate each of the 10,000 experiential variables. For no other group is the average number so low as 7. From the point of view of simplicity of description, therefore, we take the group which requires an average of 7 and designate it as *the* group of basic or primary attributes.

Presumably, if we had all of the entities within a system measured with respect to all of the experiential attributes, the principal axis method would enable us to determine the minimal number of attributes required to estimate all of the other experiential attributes. The method, however, would not tell us which of these might be most appropriately segregated into this minimal subgroup.

And this is where the second criterion is needed. Many of you have, already, doubtless anticipated a more commonly known expression for the second numerical criterion of simplicity which I have just discussed. I refer, of course, to the concept of simple structure. I have preferred, however, to formulate the concept of simple structure somewhat differently from the traditional one because I think the concept formulated in this way is less controversial.

But the crucial question is, "Will the identification of primary variables enable us to make more accurate predictions than would otherwise be possible?" Let us see how it might. In a two- or three-dimensional system it is not difficult to show that the primary variables in the system are those from which all the other variables may be estimated without the use of negative weights or coefficients. If one or more of the primary variables is replaced by a nonprimary variable in the prediction battery, then the nonprimary predictors will cause the primary predictors to take on negative weights when certain of the other variables are estimated. It is quite probable that this

principle will extend to a system of any number of dimensions, and that, therefore, if other than primary variables are included in the predictive set, negative weights may appear in the estimation of certain of the other variables. In fact, primary variables might usefully be defined as that minimal set of variables from which all others may be estimated with nonnegative weights.

The implications of primary variables for accuracy of prediction should now be more apparent. If it were possible always to find a set of primary test variables for predicting success in school or vocations or elsewhere, then, presumably, we should never have negative regression weights. For certain special cases it is easy to show that predicted scores involving negative regression weights are less reliable than predicted scores all of whose regression weights are positive. It should probably not be difficult, therefore, to set up rather general conditions under which estimates made from primary variables would be more reliable than those made from nonprimary variables. Other things being equal, then, the use of primary variables in predictive batteries should enable us to make more reliable predictions.

Is there any other way in which factor analysis might enable us to make more accurate predictions? It is easy to show that a factor analysis does not yield more information than is contained in a matrix of measures on which it is based. If the factor analysis is carried out so as to include all the information given in the correlation matrix, even errors of measurement, then multiple factor techniques yield results identical with, and hence no better than, the multiple regression methods. Most factor analysis procedures, however, assume errors of measurement. The problem is to estimate the original measures in terms of a much smaller set with sufficient accuracy so that the remaining variance may be considered as due to chance. If we assume that errors of measurement result in errors in regression coefficients, then one of the sources of error in applying regression weights to a new sample could be eliminated if errors of measurement, in the original sample, were excluded. The factor techniques may enable us, for a given sample, to get a more accurate estimate of the true correlations for that sample than is given by the experimental correlations. Therefore, it is conceivable that by means of the factor techniques we could, for any given sample, obtain a more accurate estimate of the true regression coefficients for that particular sample. The true regression weights for a given sample should, in general, be closer to the true population regression weights than would the regression weights incorporating errors of measurement in the sample. This suggests that regression weights obtained by factor techniques on a particular sample might yield more accurate predictions on subsequent samples. Therefore, the factor techniques should result in more accurate prediction of all phenomena, human, subhuman, and physical, where the fundamental vari-

ables have not been clearly isolated and agreed upon.

Let us now consider some of the limitations of factor analysis for psychological research. One of the most serious limitations comes from the fact that factor analysis has a voracious appetite for data. If you want to come out with results of any consequence you should have 50 or 60 variables or tests on at least 500 cases. Comparable form reliability should always be incorporated in the design of a definitive factor analysis. Assuming that a single form of a test should be at least 10 minutes long, the two halves would take 20 minutes. This means that you could test only three variables an hour so that it would take 20 hours to test all 60 variables. If you multiply this by 500 people, you have 10,000 man hours of testing time just to get your basic data. I would not be inclined to take very seriously the results of any factor analysis involving psychological tests, which falls far short of 10,000 man hours of testing time. According to this criterion, very few factor studies to date can qualify as thoroughly respectable. Obviously, factor analysis is not for the lone wolf operator. It is too difficult to pick up 10,000 man hours of testing time. The collection of data for factor analysis projects, I think, will more and more have to be sponsored by large scale cooperative research enterprises.

Another limitation of factor analysis is the time required for the actual computations. The computation of the table of intercorrelations can be carried out fairly rapidly with modern computing machines. But if you are a small businessman type of researcher you cannot afford the equipment required to calculate 1,720 correlation coefficients on 500 people. But assuming you can get the data and the intercorrelations, untold man hours of labor still lie ahead before the simple structure matrix is obtained.

Before starting the factor analysis, you'll have to decide what to do about the diagonal elements. Here you will run head-on into another rather serious limitation of current factor techniques. To date there is no clear agreement as to what should be used as the diagonal elements. Should you use unity? Should you use reliabilty coefficients? Should you use estimates of the communalities? If so, what are communalities? The communality of a test is a very obstreperous sort of thing that may jump around unpredictably from one test battery to another.

But even assuming that the communality is a fixed value for a given test battery, how will you define it? Can you say that the communality shall be such that the rank of the matrix is a minimum? Hardly that, because mathematically the rank of a matrix with unknown diagonals is determined solely by the number of variables in the matrix. If you define the communalities in this way, some of them could conceivably be greater than unity and some of them might be negative. You may therefore insist that in addition to determining the value of the communalities solely on the basis of the number of variables,

no communality be less than zero or greater than unity. But assigning upper and lower bounds for the communalities is a long cry from assigning specific values to them.

Assuming, however, that you could solve for values of the communalities which were within the acceptable bounds and which enabled you to account completely for the intercorrelations with minimum rank, you have violated the basic definition of communality. The correlation coefficients include errors of measurement and if you completely account for them, then the communalities must also have error variance in them. You might avoid this inconsistency by assuming that the rank s of the matrix is smaller than the rank t determined by unknown diagonals. What values now must you put in the diagonals so that if you apply a principal axis solution carried through s components, the sum of the squares of the residual r's will be a minimum. You might in turn let s take all values from 1 through t and determine t sets of diagonal values, one for each assumed rank, such that in each case the sum of the squares of the residuals would be a minimum for each rank. Assuming you had appropriate method for determining the smallest rank for which the sum of the squares of the residuals was due to chance, you might take these corresponding values as the best estimates of the communalities. This seems like a pretty respectable operational definition of a communality, but how to go about finding values which will conform to it might be very difficult.

What, now, are some of the more common arguments for the use of estimates of communality in the diagonals? In the first place, it is argued that the table of intercorrelations can be more completely accounted for with a smaller number of factors. This statement is certainly true, but it expresses a purely mathematical artifact. By the same token, the use of communalities in the diagonal leaves a greater proportion of the total test variance unaccounted for. Actually, what we have when we use communality in the diagonals is a more—rather than a less—complicated system for describing experiential phenomena. We start out with n attributes and wind up with $n-s$ attributes where n is the number of variables and s is the number of factors.

It has, of course, been pointed out that what is specificity for a test in one battery may be communality in another battery. The argument goes that if we analyze enough tests in enough different test batteries, eventually most of the specificities would be absorbed by communalities. The assumption seems to be that the use of communalities in the diagonals in a number of small overlapping battery analyses will eventually result in the same factor loadings that would be obtained if all conceivable tests were thrown into one gigantic battery and factored with unity in the diagonal. So far, however, the validity of the assumption has not been demonstrated.

49

It seems to me that the logical way to analyze this colossal matrix would be to put unity in the diagonals and fervently pray that the number of factors required to account for 90% of the total systematic variance would be a very small fraction of all the variables in the matrix. If our prayers were not answered, then we should have no scientifically honorable recourse but to commit hari kari, because the basic assumptions of scientific methodology would have been demonstrated as untenable. If, however, we decide to analyze this giant matrix by estimating the communality residuals at each stage, I suspect that the off-diagonal residuals would approach chance values when the common factor variance was still far short of 90% of the total systematic battery variance. If we found that most of the tests still had specific variance left, I would lay the results not to the inherent chaos of nature, but to the use of communalities in the diagonals.

Aside from the fact that communalities are said to result in a smaller number of factors for any given study, a more plausible justification for their use is also presented. It has been alleged that simple structure is easier to attain when communalities are used in the diagonal rather than unity. But to date no quantitative or objective criteria for simple structure have been advanced. Therefore, the question of which of two sets of rotations on the same data comes closer to simple structure cannot be answered at the present time, except to the satisfaction of a particular experimenter. Therefore, since no techniques are available for testing the claim that the use of communalities in the diagonals enables one to obtain less ambiguous simple structures, one can neither affirm nor deny the claim.

Actually, it seems to me that far too much emphasis has been placed on communality and not enough on specificity. From the point of view of prediction, the usefulness of a group of variables varies inversely as the common factor variance in the battery. Ideally, we should have a battery of measures with very low or near zero intercorrelations so that there would be no common factor variance, and most systematic variance would be specific.

There are, then, at least four unanswered questions concerning the communality. First, can you define it; second, can you calculate it; third, should you use it at all; and fourth, how should you use it?

I have suggested that I regard the concept of simple structure as a basic scientific contribution. For this reason, I think that the techniques for achieving simple structure are of great importance and that any defects in these techniques are serious limitations to the uses of factor analysis in psychological or other scientific research. One of the questions which arises in connection with rotation to simple structure is whether orthogonal or oblique rotation should be employed. We know that matrices of intercorrelations can be made to vary greatly by systematic selection with reference to all or some of the variables. It therefore seems plausible that intercorrelations of either

experiential or primary variables may well be regarded as a function of the particular sample on which the measures are drawn. Therefore, it would seem plausible that a rotation to simple structure on one group might be nearly orthogonal, while on a specially selected group it might be clearly oblique. If, then, we regard the specific character of the transformation as being a function of the particular sample, it would be unrealistic to expect that, for all groups, the transformation for any particular battery should be orthogonal. Therefore, it seems to me that the question of whether to use oblique or orthogonal transformations is no longer a critical issue.

A much more serious problem arises in connection with the actual techniques of rotation and the criteria for simple structure. At the present time, no completely objective and unique method is available for the rotation operations. One of the most pressing needs for research in factor analysis, and I believe we might say for psychology in general, is research in more adequate methods for the transformation of arbitrary factor matrices. Most of the factor techniques will indicate the minimum number of attributes required, but only the rotational procedures will identify that unique set which is maximally parsimonious for the description of most significant experiential attributes in the system. Therefore, I believe that the simple structure concept, or something equivalent to it, constitutes the greatest promise of factor analysis to psychological research while at the same time currently available procedures for attaining simple structure represent the most serious limitations.

In summary, then, the factor techniques should result in more economical and accurate predictions of socially significant behavior provided the administrative, computational, and technical limitations can be overcome.

DISCUSSION

PARTICIPANTS:
PHILLIP J. RULON, WILLIAM STEPHENSON.

DR. RULON: Mr. Chairman, if this group won't accuse me of arrogance, I should like to suggest that the first two speakers are both right. I know this will disappoint Mr. Buros, who likes to have argument. Just to show that I am not really arrogant, I will not attempt to prove that all three speakers are right.

At the Graduate School of Education at Harvard, we think we have just solved the fundamental problem of guidance, that is, we have extended the Fisher discriminant function concept to four dimensions, we think. That means—well, let me give you the example we think we have worked out, using some data provided by Dr. Henry Dyer of the Arts College of Harvard.

We have given nine tests to five groups of people, and then, extending the Fisher discriminant function concept to four dimensions—let's see; in the ordinary discriminant function, you have two groups in one dimension, you would have three groups in two dimensions on a plane, and five groups in four dimensions. We think we have computed four sets of discriminant coefficients so that, if a youngster comes along and we give him these nine tests, we can compute the coordinates for this youngster and assign him a point in space. We can then compute the cross-dimensional distance, or the slant distance, or the diagonal distance in four dimensions for this individual from each of the five groups; and, of course, the shortest distance tells which group he belongs to, just as in the ordinary discriminant function the shortest distance along the line, either plus or minus, tells which it is. I believe it was which tribe or species the specimen belonged to in the original Fisher discriminant function.

It also means, therefore, that, if this system works, it will be possible to give one hundred tests to twenty professional groups—doctors, lawyers; you know the list—and when there comes along a subject for vocational guidance, we compute his coefficients, assign him a point in space, and list the slant distances of his point from the centroids of the groups; the shorter distance, of course, determining the group he looks like.

It looks as though the system is fail-safe in the same sense that the initial discriminant function is fail-safe; that is, if we are not making any discrimination, these slant distances will all come out the same, and the

system therefore apparently has the virtue that multiple regression has—that, if you are not doing any predicting, the system says so.

Now, obviously, what Mr. Horst says about administrative difficulties in calculation are going to plague us. A tremendous amount of calculation is necessary. It took us approximately three months to get four columns of nine coefficients each in the four-dimensional case for five groups in only nine tests, and everybody knows that this work goes up approximately by the cube—I think that is correct—of the number of entries. But the reason why I would like to suggest that both of the first two speakers are right is that the approach is strictly practical, following Dr. Bennett. The validities and distances are in terms of the observed test scores themselves; that is, we collect the scores from the individual, and those are what get the discriminant weights.

But as soon as you decide to apply this thing practically—and that we have no worksheets for—you see that you will have to minimize the number of tests you give; that, if you are going to cover the distinctions between professional groups, say, with any feasible battery, you will have to use a battery which covers them maximally with the smallest number of tests, and so you have the theoretical problem of what does distinguish these people, and you will have to apply the factor-analysis method, I should say, to derive the tests that you will use in order to get maximal coverage. But the real problem is just the same, that the man comes along and wants to know to which groups he belongs, and I think he will want to know actual life-like groups—doctors, lawyers, scholars, ditch diggers, or what not—and the counselor, or the guidance officer, will not be very much interested in abstractions like factor loadings.

I think that our system is sound, but we don't know yet. I have seen the four columns of nine coefficients, but there is an awful lot of checking to do. One material problem is what is the unit of measurement in the cross-dimensional distance?—because these distances are denominate. I mean denominate numbers, so the unit of measurement is important. I would like to suggest that we are going to need (even if the system works—and we have our fingers crossed, because if the system works, we have practically what amounts to an "A" bomb in the guidance movement) a tremendous amount of machinery to implement it.

I might say as an aside that you will notice, if you do give a hundred tests to twenty professional groups, and then you call in the IBM selective sequence computer to get these weights, you have to do that only once. The guidance counselor, when his subject comes to him, takes the scores and applies the weights, but with an ordinary desk computer or with an ordinary IBM installation. The solution of the basic problem, in numerical terms, requires the use of the IBM selective sequence computer only once,

for a certain battery. However, please notice that you could not afford to pay the $300-an-hour fee for the IBM computer up in the World Headquarters building unless you were pretty sure that you have good coverage on the distinctions between these professions. And I suggest, in order to be sure you have that pretty good coverage, you had better use the analytical procedure suggested by Dr. Eysenck. But to me, the answers we are looking for, we are looking for in terms specified by Dr. Bennett.

DR. STEPHENSON: May I suggest that, instead of analyzing these things, we compound something so we don't have to handle more than one or two scores in the end?

You see, the mistaken outlook by our friend, George, is that when you factorize and discover factors G, V, K, and so forth, you should measure these separately. The truth is that there might be a type of test which involves all of these factors, and that this is the one you should want to use. It is, in fact, the one that he actually used to start with, and that is the one we might well continue to use. There is no one, certainly not myself, who thinks that we should try to measure the little V and make any particular determination from that. However, the point is that perhaps we should begin tc turn ourselves upside down and, instead of analyzing, let us pile the statistics together. Perhaps we won't need all these awful machines.

PANEL III

Information which should be provided by Test Publishers and Testing Agencies on the Validity and Use of Their Tests

Information which should be provided by Test Publishers and Testing Agencies on the Validity and Use of Their Tests

HERBERT CONRAD

APTITUDE AND INTELLIGENCE TESTS

As ALL OF YOU KNOW, the information which test publishers provide in connection with their tests generally leaves something to be desired. Sometimes we find a test manual consisting of a page, perhaps, and no more. Sometimes we find a statement to the effect that the validity of the test is assured by the care exercised in its construction ("standard techniques" having been employed), or, again, we may find merely some bland reassurance that the best evidence of the validity of the test will be found in its successful use.

You are all familiar with the excessive claims which, if not made, are insinuated. In the case of one well-known test, for example, a footnote explains that the validity and reliability coefficients presented refer to the test as of 1941, when it was a good deal longer and unabbreviated. That is in an inconspicuous footnote. Is it fair to assume that the new figures might be a little lower?

But, mostly, there is just plain nothing about what you want to know concerning the test. Nothing is said, or very little is said, or only very vague statements are given regarding the nature of the sample, the reliability of part scores, the factors measured by the test, the nature of the criterion groups, the nature of the criterion, correlations of subtests with the criterion, and so on.

If that is the case, we might justifiably ask just what ought the test publishers to provide. All of us, I think, have pretty good ideas of what is needed, but perhaps a reorganization or restatement will be in order, and at the conclusion, I shall give my ideas of how we can get these things into the reports of the publishers.

It seems to me that the test manual, or the information given by test publishers, should answer certain questions.

First, what are the purposes of the test? What purposes, what uses, are claimed for the test? That ought to be spelled out in considerable detail. If, for example, the Wechsler-Bellevue aims not only to provide a language score and a performance score, but also an indication of personality disorders, why, that should be clearly stated as one of the purposes.

Second, we should know the answer to the question, What criteria were employed in validating the test? And do these criteria match the purposes which are announced for the test? The criteria may be immediate criteria; for example, an algebra aptitude test is given by a teacher who wants to know what the achievement scores or grades will be of the students in algebra that very next semester. The criteria may be more or less intermediate, such as in the case of an elementary test being used to predict the kind of course in high school that might advisedly be taken by a student —such as classical, college preparatory, commercial, general, and so on. Or the criterion may be remote and ultimate, in the sense that it measures performance on the job as an adult. Educators, I think, are interested in all three of those kinds of criteria, and very few tests give information about even one.

Once the criteria are announced, we would have to examine into their validity. For instance, in the case of the algebra teacher, we would want to know how valid are the grades of that teacher. In the case of the high school course, we would similarly want to inquire into the validity of the criteria there, and, of course, most of all, we would want to inquire into the validity of the final or ultimate performance-criteria.

So far, we have not mentioned anything about the tests except their purposes. We have talked about criteria. With regard to the tests, there are a great many questions.

First of all, how much of the test score represents nothing more than chance, that is to say, how much of the test is, so to speak, nontest? That question applies not only to the test itself but also to part-scores or subtests.

Secondly, we should like to know something of the internal consistency or functional unity of each of the sub-tests. That is usually obtained by item analysis. Relatively few tests report data on item analysis in any detail.

Third, we should like to know to what extent the test measures a mixture of speed and power. At the present time, so far as I know, there is no very uniform method by which the speed component is measured. Some people use the number of cases or proportion of cases attempting the last item. Dr. Tucker of the Educational Testing Service uses, among other things, the ratio of the standard deviation of the number of items answered correctly to the standard deviation of the number of unattempted items. Obviously, if the standard deviation of the unattempted items is very great, then the speed factor is, presumably, fairly important in determining individual differences in the test scores.

Fourth, we should like to know how this particular test compares with other tests. How is it related to other tests? Here, what we need is a table of intercorrelations—which seldom graces a test manual—or, beyond that, a factor analysis, with the particular test in which you are interested placed

in a battery that is especially designed to reveal the factorial composition of the test.

Fifth, we should like to know to what extent the test is affected by various external or extrinsic factors, such as practice, coaching, special experience, cultural factors, and so on. This applies particularly to tests of aptitude and intelligence, which are the subject of this talk. Especially in the case of mathematical aptitude at the higher levels, it seems very difficult to obtain a measure of aptitude that will not reflect a person's courses in mathematics—their recency, the grades that he made in them, his diligence, and so on.

Sixth, we should like, I think, to place this test into a correlation matrix where the other members represent criteria, or criterion factors, since one of the prime requirements of a test for greatest usefulness is that the test correlate high with one criterion and low with the others. That is essential if you are to have any kind of differential prediction and, as was pointed out by Dr. Rulon this morning, the problem of guidance is basically one of differential prediction.

Seventh, we ought to know what is the contribution of this test over and beyond what is available from other, easier sources. For example, it is very easy to find out the person's chronological age; will our measure of aptitude tell us something that chronological age does not already tell us? It is also easy, in some cases, in a local community, to obtain the person's school record. If that is so, then the question is, Does the intelligence test or the measure of aptitude tell us anything that is not already told by the previous information? The *independent contribution* is certainly something which should definitely be known, but very seldom is it revealed to us in the information which test publishers provide.

Eighth, we should ask ourselves, What is the effect of the test on the person who takes it? Most of these aptitude tests are taken in the environment of the school. Does the giving of the test leave the individual discouraged, feeling more hopeless? Does it lead him to think that school is where you are given questions you cannot answer? In other words, what are the side effects, so to speak, of the examination?

Some tests have scoring arrangements—I am talking now of certain tests of the Educational Testing Service—such that either norms are not immediately available, or the scores do not become immediately available. This is a serious handicap, so far as use of the test for practical guidance is concerned. I was talking last summer with Dr. Frank Fletcher, Head of the Occupational Opportunities Service at Ohio State University. For their purposes, promptness of scoring is virtually essential. They bring in a person, let's say, from Cleveland, Sandusky, or other parts of Ohio, for two days of testing and counseling. They might give the Educational Testing Service's pre-law test, but since the test is scored back at Princeton, they would have no satisfactorily

prompt knowledge of what the law test scores are. Now, there may be administrative problems, problems of expense and other problems, that produce these late norms and late scores, but so long as those problems remain unsolved, the usefulness of the test in a guidance situation such as that described will inherently be curtailed. It may be that the prime purpose, the original purpose, of the test is fulfilled, but it seems too bad that other normal purposes cannot also be fulfilled.

And, finally, it seems to me that we ought to expand our notion of validity and responsibility to include, let us say, the elimination of muddleheadedness by the users of the test. I can give you a clear example, close to home. Colleges and universities require that students applying for admission to college take the Scholastic Aptitude Test, giving a Verbal and a Mathematical score, and, in addition, take an English Composition Test. It has been found in a study by Miss Edith Huddleston, which I think is a model in many ways, that the Scholastic Aptitude Test will predict English composition scores and grades better and more cheaply than the English Composition Test. This was discovered at least one year ago. But unless colleges are much more responsive than I think, I dare say that these colleges are still requiring the Scholastic Aptitude Test and the English Composition Test.

I think it is the responsibility of the test publisher to point out very vigorously to such persons that so far as can be seen, the giving of the second test, the English Composition Test, is not necessary. It is a waste of money and a waste of time. In the time required by the English Composition Test, some other test could be substituted, for other purposes and to better effect. What I am saying, in effect, is that matters of policy are not, or should not be, considered outside the province of the test publisher. Let me give an analogy.

A person who is producing drugs, an ethical manufacturer of drugs, does not simply put a drug on the market and say, "Well, if it is misused, it is not my fault. After all, it says here in the print that they don't have to misuse it if they don't want to and if they have any brains." The ethical manufacturer goes to some little trouble to see that the thing is not misused. The same is true, let us say, of the manufacturer of special equipment. New radar equipment, set up on a ship, would not be worth very much unless the manufacturer went on that ship and saw that the radar equipment was used properly. That is part of the manufacturer's job. It should be part of the test publishers' job to take the same responsibility.

Well, I dare say that all of this is more or less familiar to you. We know that at least *most* of these things—I think you will agree—should be done. What I should like to emphasize is that *we ought to follow through* on our testing, and not simply say that "Here is available—." If we think

that local norms would be useful, we ought to do what we can to see that the local norms are correctly obtained. If we feel that just giving the means and standard deviations is not enough, we might give an expectancy table and indicate how it may be used; with that expectancy table, we ought to be very careful to point out that the expectancy for extreme scores is less reliable than the expectancy, let us say, for scores around the mean.

But, to go on, granted that this information is desirable, or that most of it is desirable, we know it is not forthcoming in most cases. Why is that? Is it unnecessary? Most of it is both desirable and necessary. Would giving the information possibly reveal too much of the shortcomings of the test? I think that is true, in some cases, but no ethical publisher would on that account fail to reveal the facts about his test. If his test is defective, you can be pretty sure that other tests are equally, if not more defective—assuming that the publisher is experienced and capable in the selection of tests which he publishes.

Is it too expensive to provide the information we have asked for? The answer to that, I think, is a resounding, "Yes, it *is* too expensive." I can imagine that, if the Educational Testing Service, for example, did half the things I have been talking about, it just could not continue in business without continual grants from outside sources. In other words, an ethical producer of tests, if he is asked to meet all of these ideal requirements, would soon find that he is the goose that tries to lay the golden eggs with every test, but there just isn't enough of that gold in the system.

What can we do about the matter? Well, we can't pass a law. That seems pretty clear. Can we educate consumers to demand and pay for the superior product? I think that college teachers are succeeding in educating the consumer; the test producers are also educating the consumer. It seems to be a pretty slow process. It is basic. It is necessary. Is there anything that could be done to speed it up?

I have one suggestion which I think has been made before, namely, the creation of an impartial bureau of standards for test validation, which would have enough prestige so that its word would mean something in the market of tests. Such a bureau should not be connected with any university. It is "hot" business, this business of validating tests and saying which is which, which is superior or inferior, and so on. A university in general cannot stand it. And it should not be a governmental enterprise, for the same reason. Nor should it be an enterprise dominated by any one test-producing organization, because if it were, the results would be suspect. I feel sure that if the author of the Ohio State Psychological Test, for example, were to validate his test, while somebody else in the Educational Testing Service validated the Scholastic Aptitude Test, it would be pretty hard to avoid unconscious bias, and there would always be the suspi-

cion that unconscious bias had crept in. So it has to be an independent organization.

Well, of course, the question comes in, Where is the money going to come from for that? The money can only come, it seems to me, if the bureau acquires enough prestige so that the consumer is willing to pay for the badge of approval and for the stock of information—willing to pay enough extra, so that this bureau can be supported on a trial basis from the test-producing organizations themselves.

Information which should be provided by Test Publishers and Testing Agencies on the Validity and Use of Their Tests

PAUL L. DRESSEL

ACHIEVEMENT TESTS

Present Practices in Providing Evidence on Test Validity

ALTHOUGH the emphasis in the title of this paper is on "should be," it seemed to me advisable to look first at what "is" as a basis for talking about what should be. Perusal of manuals, supplemented by the reading of reviews in the *Mental Measurements Yearbooks*, resulted in the listing of many different types of evidence which have been provided on the validity of achievement tests. These can be classified into five relatively distinct types of evidence, as follows:

1. No evidence
2. Expert opinion
3. Current practice
4. Statistical
5. Face validity (Validity by assumption)

The *no evidence* category means just that. The author and publishers furnish no evidence on validity, on the criteria for construction, or on the significance of the results. *Expert opinion* includes statements to the effect that validity or appropriateness of content are insured by the experience of the authors, the criticism of specialists, or adherence to the recommendations of certain councils or committees. *Current practice* includes statements that the test content represents best teaching practice, that it is based on reputable criteria or an analysis of textbooks. *Statistical evidence* includes data or statements indicating that the test discriminates between good and poor students, that the scores correlate to such and such an extent with grades or point average, that all items used have been statistically validated, that items have been arranged according to difficulty. *Face validity* is a much abused and, at this stage, somewhat disreputable term. Obviously, statements which I have classified as expert opinion and current practice might be considered evidence of face validity. In terms of Mosier's classification of types of face validity, these would seem to involve validity by definition or on the basis of previous research by others. There are, again according to Mosier, two additional types of face validity: The assumption of validity on the

62

basis of a common sense relationship to the objective, and validity by appearance. When the simple statement is made that the test has face validity, the author or publisher seems most frequently to mean one of these last two—that is, validity by assumption or by appearance.

Use of expert opinion, current practice and face validity as evidences of validity is questionable, particularly when as in many, perhaps most, cases, the experts, the basis of determination of best teaching practice, the textbooks, etc., are not made known to the test user. The user has no way in which to check the statements made and has the alternative of accepting them at their face value or of ascertaining validity through his own efforts. Such reporting is not in accord with the principles of scientific research.

Relationship of Validity to Reliability

The review of test manuals also revealed the following tendencies regarding validity and reliability:

1. Tests of high reliability are frequently assumed to be highly valid.

2. Items yielding high reliabilities are consistently selected over those yielding low reliabilities.

3. The number of objectives measured is restricted for the sake of greater homogeneity and consequent higher reliability.

4. No distinction is made among the various sources of variation, such as:

(a) Variations among or within individuals

(b) Degree of difficulty of the material

(c) Sampling of the area

These tendencies suggest that despite the extensive amount of material written about validity and reliability, there exists none too clear a distinction in practice. This confusion is compounded of ignorance, lack of clarity in the concepts, and also lack of appropriateness to the basic purpose to be served. For example, various meanings assigned to validity include:

(1) The extent to which a testing technique gives evidence of mastery of the desired technique.

(2) The extent to which the test indicates status relative to the universe of which these items are a sample.

(3) The extent to which we can predict something from the test.

(4) The extent to which the test indicates the ability to handle real life situations—that is, situations outside of the classroom.

These four concepts, except for some rewording, are essentially ones mentioned by John Flanagan at the 1948 session of this Conference. Any one of them is applicable to achievement testing in some ways and yet also inapplicable or at least unsatisfactory in others. Let us examine each concept to clarify this.

Recalling a fact and selecting it from a group of proffered responses are not the same. However, a completion test and a multiple choice test on the same facts can be set up, and the correlation between the two is an indication of the validity of the multiple choice test as a substitute for

the completion test. However, as one deals with more complex objectives the problem of accurately rating an unstructured response becomes so great that the decision as to whether a structured test situation indicates mastery of the desired objective becomes more a matter of expert judgment than of statistical manipulation. If our objective of recall of facts becomes one of recall at appropriate time and place for use, we face a difficult situation.

The second concept of validity—the indication of status relative to the universe of which these items are a sample—is productive of confusion with reliability. If the items involve skills somewhat different from the desired skills, the measure of the extent of overlapping or similarity of the test items and the desired skills is an indication of test validity. The relationship of a sample of the slightly invalid items found in the test to the universe of all such invalid items may be more properly thought of as the test reliability than as its validity. On this basis, the correlation of a test with a longer test of the same type is not ordinarily a measure of validity. It will be only when the test items are shown to be valid in the first sense—and then validity and reliability are identical.

Relative to the third concept of validity—prediction—the difficulty lies with the criterion. If we are to accept the usual course grades as criteria and construct tests so as to yield the highest possible correlations with such criteria, we shall scarcely improve either the tests or educational practice. A well constructed test can much more logically be used to investigate the validity of usual grading practice. We have, in fact, validated tests with grades and grades with tests rather indiscriminately, to the confusion of everyone.

A student last year asked me to explain just this matter. He said, "The Dean says objective tests are used because they are more accurate than instructors' grades. The instructors say the tests are no good because we don't get the same grades. The State News reported that a study had shown a high relation between tests scores and grades. I'm confused."

Validity conceived of in terms of the extent to which a test provides evidence of ability to handle real life situations puts the maker of a single test in an almost hopeless situation insofar as producing any statistical or other evidence of validity. The adequacy of performance in a real life situation is a judgmental matter, and it is entirely too complicated to pick out that which is relevant to a particular test. Moreover, planned situations are not real life situations, so that it becomes necessary to make extensive and surreptitious observations on all individuals involved in the validity study—a program apt to be productive of situations in its own right. Conceivably this sort of thing may be done to study a total program. It can hardly be done for a single test.

We appear to be eliminating all basis for establishing validity and we are not yet finished. For an aptitude test which measures some trait sup-

posedly not much influenced by training, validity may be construed as the relationship between the test performance and actual performance. An achievement test, however, is used in connection with objectives supposedly and, at least occasionally, actually influenced by education. We are then faced with these additional difficulties:

(1) Validity does not exist except in relation to a particular individual in a particular educational program. This is exemplified by the fact that

(a) an item constructed to measure reasoning does not do so if it has been taught in the course and is then recalled.

(b) an item to measure reasoning is not valid for the single individual who may have seen that particular problem somewhere, even if it is for the rest of the students.

(c) an item involving an objective not accepted or not developed for a course is not valid for that course.

(d) a test not involving all the objectives of a course is not valid unless this deficiency is recognized.

(2) In an attempt to get validity, items are selected on the basis of evidence that students show change with regard to them. Our tests are then loaded with materials on which change is most easily obtained and tend, therefore, to perpetuate certain (and frequently bad) instructional practice. Test items which are best in terms of statistical evidence are frequently trivial.

(3) Emphasis on national, regional, or even local norms is frequently related to validity in that such norms show the wide differences in performance of individuals and of groups. Actually this evidence tends to obscure the more important questions of

(a) actual gains by individuals and groups relative to various objectives

(b) what constitutes a reasonable or an optimum gain over a given period or as a result of certain courses

(c) differential gains for students of differing abilities. The most capable students should show the greatest gains—not just be at the top of the distribution.

Up to this point we have been critical and destructively so, but we can now base a positive and, I hope, constructive set of recommendations on this criticism. At the first stage it would seem to be necessary that some attention be given to clarification and careful statement of some of the more important educational objectives (not necessarily only those most emphasized in practice). These should be defined not just in words but in terms of actual problem situations, with detailed analyses of the kind of behavior which is or should be elicited.

Based upon such a statement we would then expect a test author or distributing agency to

(1) state the specific objectives covered by a test and indicate the relative emphasis on each.

(2) provide evidence that the think-

ing done by students in handling the tests is that involved in these objectives.

(3) provide evidence of the initial status of various types of groups relative to these objectives.

(4) provide evidence on possible and desirable gains which can be made under certain conditions.

(5) list the knowledge, facts, and principles needed for handling the test. This is particularly important for tests involving applications and critical thinking.

(6) list desirable educational objectives not covered by the test and suggest ways of supplementing to get a well-rounded evaluation program.

(7) de-emphasize status norms and place emphasis on growth.

(8) describe the level of functioning of individuals falling at various points in the distribution of test scores.

(9) indicate the extent to which the sampling of a particular type of behavior is adequate for ranking individuals with regard to that type of behavior.

(10) provide detailed information on textbook analyses, experts, instructional practice, or other such expert opinion or current practice approaches to validity.

It may be objected that these suggestions do not result in neat statistical summarization of validity and reliability. I, for one, don't care. If such a program were followed the prospective test user could select and use a test with some understanding and with some assurance that the results would be useful. He would be encouraged to pre- and post-test to determine the amount of actual progress made. He would have some idea of what to do if he found that inadequate gains were registered. Such tests would be valid because they would have value and they would be reliable because teachers would rely on them.

I really have said what I have to say on the assigned topic, but I have one more related point to make. Test makers have been in a state of unstable equilibrium in that they have been constantly reaching ahead for opportunities to show teachers the value of scientific testing and at the same time leaning over backward to avoid the criticism of controlling or determining curriculum or teaching methods. In trying to mold tests to fit teaching practice, testing has been guilty of perpetuating practice. We cannot avoid influencing instruction, and we had better face the issue clearly and boldly.

I believe that a qualified subject-matter man who seriously turns his attention to evaluation attains a degree of insight into the student mind and into the significance and inter-relationships of the subject that is far beyond that of the average teacher.

I believe that as such an evaluator seeks for and finds situations which give opportunity for a student to show whether he has a certain skill or ability, he also finds that these same kinds of situations are the best situations in which to place the student so that he may get practice in such a skill or ability.

I believe that the evaluator has an obligation to show how such ma-

terials can be used to improve instruction and that he should be more concerned with this than with preserving the secrecy of his test items so that they can be used for grading.

I believe that many instructors, particularly in general education programs, are sincerely trying to find ways of developing thinking skills and other objectives beyond factual knowledge without quite knowing how to do it, and if we do not adapt our evaluative materials for instructional use we play the fool by trying to test and grade students on something for which they have had no training.

Our first job in achievement testing is to assist teachers to get maximum achievement. Ranking or grading should be incidental and yet I believe it is obvious that to date our testing has served primarily to concentrate the attention of teachers and administrators on grading, ranking and selection. It is high time that measurement and evaluation assist rather than distract the classroom teacher.

Information which should be provided by Test Publishers and Testing Agencies on the Validity and Use of Their Tests

LAURANCE F. SHAFFER

PERSONALITY TESTS

THIS is an era of personality tests. Hardly a month passes without bringing some new measure of personality, either issued by a test publisher, or announced in a book or journal article. The appearance of these many tests is itself a social phenomenon of great interest. Educators and psychologists are no longer content to deal only with the intellectual and academic aspects of people, but increasingly recognize the importance of motivation and emotion in human life. Mental hygiene and personal adjustment have come out of the clinics, and are emphasized in education, industry and many other fields. The flood of personality measures in a response to a great need to do something constructive for the welfare and effectiveness of whole people.

Although personality tests emerge from a great and commendable need, most critical observers find that their quality is less than their sheer quantity. Few if any have firmly established validities for the purposes for which they are being used. Serious difficulties confront any worker who attempts to validate a measure of any aspect or aspects of personality. The need for the tests is so great, however, that many come to the market sadly lacking in evidence of their worth. The basic equation seems to be that a pressing demand for personality measures, plus almost insuperable technical difficulties, equals many bad tests.

What data should be supplied to attest to the validity of a personality test? A first and very modest answer might be that *some* evidence should be given rather than none at all. Even so small a demand is unmet by many tests. During the past year two elaborate clinical tests that draw sweeping conclusions on many aspects of personality have been described by whole books. In each case, the only evidence of validity is an appeal to the clinical experience of the author, derived from studying hundreds of cases. The user is asked to try the test on faith, and to seek evidence of validity through his own experience with it. Such an appeal to private and subjective feelings of validity, in place of its public demonstration, seems more

likely to lead to mystic cults than to scientific progress. These two tests may be very effective measures of personality indeed, but we shall never know with certainty until they are subjected to sound and communicable investigation.

Tests whose simpler design makes them easier to study objectively have also been published without validation. There are numerous questionnaires intended to select students whose personal maladjustment demands individual attention, which are quite unaccompanied by any evidence of having identified such students. The items and scored responses have been selected because experts think them pathological, or because they have been included in preceding similar questionnaires, or because they correlate with one another in internal consistency. Such instances may lead test users to read manuals critically, and to refrain from using totally unvalidated tests except for research studies.

Although all tests should possess some evidence of validity, the nature of the information may vary greatly according to the character of the test. Not all personality measures are alike. They differ widely both as to the manner in which they are applied and the uses for which they are designed. Some of the problems of validation can be clarified by considering the nature of the tests.

Personality tests vary in the degree to which they depend on the skill of the examiner. At one extreme are clinical tests from which no scores are obtained, that are evaluated qualitatively as symptomatic of the personality of the examinee. Examples of techniques at this end of the continuum are the Thematic Apperception Test and other picture-story fantasy tests. No novice can use such a test effectively, for it is only a means for eliciting responses that are evaluated by a thorough knowledge of the dynamics of human behavior. The fundamental measuring instrument is the clinician himself rather than the test. When the test has no validity apart from its user, it is perhaps irrelevant to speak of its own validity. One can, however, study the degree to which the clinician is valid with and without the aid of a certain technique, and thereby assess the value of the test indirectly. Up to the present, few such studies have been reported.

At the opposite pole of the same continuum are questionnaires that are administered objectively and scored by machine or even by the examinee himself. Here dependence on clinical skill is certainly at its minimum, and the test instrument can be evaluated in isolation. Therefore, the test maker has a clearer obligation, as well as a somewhat easier opportunity, to report validity in unambiguous terms.

Another important way that personality tests differ is in respect to the number of hypotheses that they undertake to explore. A single-hypothesis test is one yielding a single score, which undertakes to predict a single outcome. Since that situation gives the best opportunity for explicit validation, the demand for evidence should be most rigorous. In some instances, it is practicable to set up a validation

experiment with a design like that used to study the effectiveness of an aptitude test. Several questionnaires constructed in the armed forces during World War II were reasonably valid predictors of failure to adjust to the personal requirements of military service. Perhaps the unusual motivations of men in war time contributed to the success of such tests. Equally important, and still insufficiently explored in analogous civilian situations, were the decisive and soon-available criteria, the large numbers of cases tested, and the rigorous analyses of item validities.

The use of operational criteria could resolve many of the existing difficulties in the validation of personality measures. It is hard to validate a test that claims to identify "neurotic" students, because no one can define exactly what is meant by "neurotic." Tests to identify students who will seek the help of a counselor, or who will get lower grades than indicated by their abilities, or who will drop from school for other than academic or economic reasons, might be practicable because the criteria are functional ones.

When a test has been constructed to predict a specific outcome, the user wants to know in simple terms how well it will function. In addition to the usual validity data expressed in terms of correlation or of the significance of the difference between groups, the test maker can give some very meaningful percents. In a straightforward prediction, the percent of correctly identified cases, of false positives, and of false negatives will be of greater value to the user than will many more sophisticated tables.

Many personality tests attempt to investigate multiple hypotheses rather than single ones. Their aim is to describe the whole personality, or large areas of it, rather than to answer specific questions. Examples of these tests vary from questionnaires that yield profiles of from four to ten scores, to clinical devices such as the Rorschach from which almost infinitely varied personality descriptions are drawn.

Multi-dimensional personality tests offer problems of validation that have never been solved satisfactorily. Indeed, it is possible that the ordinary concept of validity is completely inapplicable to them when they are considered as broad descriptive methods. A full description of personality can be validated only by comparing it with some equally full description obtained by another method. The question then arises as to what is the criterion and what is the dependent variable. If we compare the whole Rorschach with the result of a long and well planned series of interviews, which is the criterion? Do we judge the Rorschach by the interviews, or the interviews by the Rorschach? It is feasible and valuable to compare the techniques, but the comparison is not a validation in the ordinary sense.

The user of a multi-dimensional test often conceives his purpose as purely descriptive; he wants to understand the person fully and deeply. In many instances that is a worthwhile objective. To really understand an-

other person, for the sake of the understanding itself and without trace of ulterior motive, is a rich and satisfying experience, akin to the aesthetic appreciation of great art or great music. In other cases, the source of the need for understanding is not so pure. To pry into another person's inner life, to crack apart his defenses and view his basic motivations and conflicts, may be a peeping-Tom kind of satisfaction, to serve the curiosity of the examiner and his need for power. In any instance, aesthetic, morbid or the many degrees that lie between, purely descriptive personality study for its own sake alone cannot be validated; it can only be experienced.

Users of multi-dimensional personality tests do not always operate on the abstract level of pure description. They have practical questions to answer. The Rorschach worker asks his test whether the patient is schizophrenic. The college counselor may consult a profile drawn from a questionnaire for a clue as to whether the student's anxiety is centered on academic, family, or peer-relationship problems. These become validatable issues. When a complex test is applied to a specific problem, criteria can be identified, and designs constructed to study the relationships between the criteria and the test.

It is very likely that many personality tests can never be validated "as a whole" in their descriptive functions. But all applications of such tests are subject to the scrutiny of validation. The maker of a test therefore has a responsibility to determine the uses to which a test will be put, and to supply dependable evidence relevant to such uses.

In stressing the functional definition of validity, the emphasis so far has been on the question of validity for *what*. Equally important is the need to define validity for *whom*. Although the culture of the examinee has implications for the interpretation of all tests, it has special significance for personality measures. One does not have to go to Bali for examples. Attitudes toward authority or toward sex that would predict serious personal maladjustment in a boy in a small suburban town may be quite normal for a youngster in a city slum. All too many generalizations substantiated only on mental hospital patients are being extended to college students and other normal adults in the interpretation of some personality tests. It is incumbent on the test maker to specify in some detail the population for which his validities hold. The day is past when "300 ninth grade pupils" can be regarded as an adequate description of a validation group.

In summary, the validation of personality tests should keep close to operational reality. A personality test as such can hardly be validated. What can be validated is the use of the test for a particular purpose, and with respect to a defined group of examinees. The maker and publisher of a personality measure have an obligation to obtain and furnish information that will let the user know the degree of confidence with which he can apply the test to practical human needs.

DISCUSSION

PARTICIPANTS:

OSCAR K. BUROS, WALTER V. BINGHAM, WALTER N. DUROST, HAROLD G. SEASHORE, PAUL HORST, HERBERT S. CONRAD, JOSEPH ZUBIN, HERSCHEL T. MANUEL, ROGER T. LENNON, IRVING LORGE, PAUL L. DRESSEL, LAURANCE F. SHAFFER, TRUMAN L. KELLEY.

CHAIRMAN BUROS: The recent issue of *Personnel Psychology* has a column by Dr. Bingham which is both encouraging and discouraging. It is encouraging because it is the presentation of a forceful, excellent idea. It is discouraging to me because, as Dr. Bingham mentioned in his column, this idea has been presented for at least the past twenty-five years. It appears, then, that the lag in the use of expectancy tables is at least twenty-five years and possibly thirty or thirty-five. I should like Dr. Bingham to say a word about this.

(Reprints of Dr. Bingham's paper entitled "Great Expectations," from *Personnel Psychology*, Vol. 2, No. 3, Autumn, 1949, were distributed.)

DR. BINGHAM: Mr. Buros and Members of the Conference: I was reading, the day before yesterday, the report of the meeting a year ago, when you were discussing the construction of tests. A most stimulating and, to me, provocative thought was recorded there in some remarks of Mr. Langmuir who insisted that before starting to build a test we should undertake to define what we want a score on that test to tell us.

Now, he did not say, "what we want the coefficient of correlation to tell us," but, "a score." The meaning of a score is what those of us want who are concerned with counseling individuals and who have occasion to look at the cumulative record of achievement test scores, personality appraisals, and aptitude tests.

We want each individual's score to tell us something about *him*. What is it we want to know? Most of us, I believe, would be grateful to have at hand the information that would help us to know what such-and-such a score on this test tells about the *probability* that if a person who makes that score goes ahead and takes calculus, or enters the dental college, or undertakes to learn carpentry or whatever the test is for, he will or will not succeed *as well as the average member of the particular criterion group*

I am glad that preceding speakers, Dr. Conrad in particular, reminded us of the ethics of the pharmaceutical manufacturers. I have been looking into some of their publications about new drugs. They are meticulous to state the contraindications, and to publish the facts as to when you must not use this drug because not enough is known about it yet. "You may use it in such-and-such conditions," they say to the physician, "but not in these others for which it might seem to be useful." Can't we, Mr. Chairman, raise our profession to the ethical level of the pharmaceutical manufacturers and the businessmen who sell drugs to physicians?

CHAIRMAN BUROS: I was especially interested to hear Dr. Bingham compare the amount of information provided by test publishers today with fifteen years ago, to the discredit of the test publishers today. In a paper which I gave at the American Psychological Association meeting in Denver last month, I made the statement, which I will repeat at this time, that the better test publishers today are supplying less information about the construction, the validity, the use, and the interpretation of their tests than the better test publishers twenty-five years ago. The meeting is now open for discussion.

DR. DUROST: Mr. Chairman, I would be remiss in my responsibilities if I were not to rise to that challenge, having been the Director of the Test Division of World Book Company for a period of thirteen years.

I think that is quite incorrect so far as any of the publications by World Book are concerned.

DR. BINGHAM: Quite so.

CHAIRMAN BUROS: And I will add to my remarks that one of the reasons why I can make that statement is that twenty-five years ago today the World Book Company played a much more important part in test construction than it does today. I did not name the company in my talk about it at the APA meeting, but I stated that only one company has been consistently attempting to give adequate information about the construction, validation, and the uses of its tests. I think that is primarily due to the influence of Truman Kelley and Giles M. Ruch, who insisted right from the beginning that they place the cards on the table. Now, the World Book Company does not have a clean record, by all means, but compared to the other publishers it has a record they can look at with envy. Do you agree with that, Dr. Bingham?

DR. BINGHAM: I do.

DR. DUROST: Well, thank you very much. I would like to make one more point—

CHAIRMAN BUROS: Excuse me, but let me mention just one more point on that, to give you an idea. In the days when Truman Kelley and Ruch and Terman were putting out the Stanford Achievement Test, they furnished a wealth of information about those tests. Now, the current edition of the Stanford Achievement Test does not possess that wealth of information. As a matter of fact, that

manual is nonexistent. In other words, the World Book Company is slipping, I think, from the days when it had the influence of these early pioneers in the testing movement.

DR. DUROST: May I say that that manual was written and only the untimely death of Dr. Ruch prevented its publication, and that the later Metropolitan Series is accompanied by a manual of the sort you describe.

I would like to make what seems to me a much more penetrating comment on this whole problem; that is, that the test authors, after all, are the ones who should be pressuring the publishers and not the other way around. You do the publishers a great deal of honor in suggesting that they should be the ones to set the standards but, after all, historically speaking, the publishers have only been the medium by means of which the production of a professional group of this sort has reached the public, and it certainly is an odd circumstance when the publisher has to pressure the profession to give the information that the profession should give in its own right.

DR. SEASHORE: Mr. Chairman, I don't want to quibble with your statements but I would like to see the document some time. I would like to speak to a more fundamental problem.

The discussion by the panel today turned out to be pretty much a discussion of the nature and quality of validation research and only incidentally was it a discussion of what should go in the test manual. Even Dr. Conrad's rather cogent remarks, which were more directly concerned with manuals, revolved around that problem. The implication is sometimes given that the publisher is responsible for all the validation research. We like to pride ourselves as psychologists —not as publishers but as psychologists—that we have a free working world in which all users of tests are free to do research. I think it is a matter of record that for most of the tests now existing the bulk of the validation research was not done by the author and was not done at the cost of the publisher, but was done by independent, free-operating psychologists, wherever they happened to be.

I think that is good. I think it is necessary. It is independent. When a test publisher puts together a manual, his chief source of information is not the author but the hundred and one, perhaps five hundred and one, test users who will share the information with him. I can report that, for most of the manuals which I have been responsible for editing, the best validation research has not necessarily been the author's, but it has been independent research by university people, by personnel people, and by school users.

That ties in with the notion that Dr. Conrad gave a while ago, that publishers have a responsibility for servicing their tests. I think we do have a responsibility for servicing our tests, and if we could raise the prices high enough, we could have a psychologist in each package. That ought to work.

Let me give you some economics on this. A test user who purchases

about $300 worth of tests a year calls us by long distance and says, "I'm being heckled because some of your tests aren't too good." Therefore, we sent Al Wesman out, at a cost of $200, to find out what goes on in that community. The rest of you paid for that trip. The $200 cost of that trip could not come out of the $300 a year that this one school pays for that special service. We did it because we thought we had a hot spot that needed some fixing and it did.

Now, the important thing is that the test publisher cannot give that kind of service out of comic book prices for tests. If tests are to be sold on a service basis, then there must be a service fee attached, and a test publisher can go only so far in that. We do our very best by trying various methods of publicity, issuing reports, and making speeches at various places. Maybe we don't do too well. We do about as well as $40-a-week clerks like us can do in the short time we have available to do it. But that is all we can do.

I think the responsibility for the quality of tests and their usage does not reside in the publisher but in the fifty or sixty graduate schools of this country who train test users.

CHAIRMAN BUROS: I should like to let you in on another little secret. Ordinarily, I don't say nice things of the test publishers, that is, with respect to their tests or the manuals, but when the manual for the Differential Aptitude Test battery by Bennett, Seashore, and Wesman came out I was so deeply moved by the unusual information they were giving in that manual that I wrote a congratulatory note to the publishers. That is the only time that has ever happened—although I did have a little postscript in it saying, "However the main part of the section on validity, is yet to come." Nevertheless, it was, I think, a very fine, honest manual and they made the statement without any qualification that they did not have the data on validity and that they warned the test user that it would be forthcoming, and it has been coming out in looseleaf form.

Are there any other questions?

DR. HORST: I want to say, first, that I have no particular test publisher or testing organization to defend. Second, that I was very much impressed by the list of criteria that Dr. Conrad gave as necessary for qualifying a test or giving adequate information about a test. I think it is extremely important, and I think something should be done about it. I would like to get some kind of an estimate from Dr. Conrad as to how much he thinks it would cost to provide the kind of service that he suggests in his bureau. Now, I think it would be an extremely good idea if it would work, but I am going to stick my neck out and then I am going to ask Dr. Conrad to stick his neck out, because by my high-speed electronic computer, I have just calculated that to provide all the information that he wants and which he has every right to expect, an average test would cost the user, let's say, roughly two or three hundred times as much as it does at present. I

would be interested to know what he thinks it would cost.

DR. CONRAD: I probably think the same as you do, but I would make this reservation, that the second form of a test could be treated in more summary fashion than the first: that is to say, the second form would have a certain amount of validity by resemblance, let us say, to the first form. That is not ideal. There should be, perhaps, a thorough check every decade, shall we say, in the case of a repetitive test, but the cost of doing good work is certainly an inhibiting factor.

During the war, I think that some very fine work was done in the construction and validation of tests in the Air Forces, in the Adjutant General's Office, and for the Navy, most of it, or a good part of it at any rate, under the Applied Psychology Panel and in the Navy itself. I don't think that any individual organization at present could afford to do a thorough job on many tests. I do think that the consumers have to be educated to being willing to pay more for tests which are certified by some central impartial agency in which they can have faith.

DR. ZUBIN: Mr. Chairman, this plan of Dr. Conrad's seems to coincide with the plan of the Committee on Diagnostic Devices of the Clinical Section of the American Psychological Association which is attempting to do something about the plethora of personality tests that are flooding the market. Not much can be done about the tests already in existence, but the proposal of this committee is that whenever a new test arrives (and they arrive, as Dr. Shaffer said, almost one a month), ten centers be selected throughout the country which will attempt to duplicate the results obtained by the test-maker on a similar population. If the test claims are found to be valid, an unofficial stamp of approval might be given by this committee.

The field of personality tests is a little different from the field of educational achievement tests because the composition or the population dealt with very largely determines the outcome. We might expect the first two or three try-outs of this method to arrive at contrary results. In such instances I hope that we could obtain the data in the various centers and see why the test worked well in Center A and not in Center B. This could supply us with information not only about the test, but also about the comparability of patient classification in various centers.

I want to take this opportunity to comment about Professor Shaffer's statement regarding the relative values of interviews and tests. I think the skills of interviewing are in danger of vanishing in the psychological field. There seems to be a growing tendency to have less and less reliance on the interview and depend more and more on such diagnostic tools as projective tests. There has been too much dependence on ink blots and similar devices at the expense of the basic approach to understanding personality through interview. This is a good example of placing the cart be-

fore the horse. Rorschach himself, for example, based his unfinished attempt at the validity of his test on interviews with patients conducted either by himself or by the referring psychiatrist.

I think this attempt at hiding behind unvalidated tests is very unfortunate. It seems to me that we ought to recapture the kind of procedures that Colonel Bingham wrote about fifteen or twenty years ago in his book on interviewing and instead of hiding behind tests which in turn depend upon interviews, standardize the interview itself, through recordings and modern rating scale techniques. Then if the diagnostic aids are found to be useful, all the more power to them. But until the evaluation of the interview itself is standardized, there is little hope of validating the tests which depend on them for validation.

CHAIRMAN BUROS: Are there any other comments?

DR. MANUEL: I want to make a point that I think has not been made sufficiently well. We have been talking about what a publisher ought to put in his manual, as if the consumers were to be members of this group or members of other groups similarly trained. Well, that just isn't true. That is unrealistic.

We need something more than the things that have been emphasized to this point. It is not hard to get such validity data and reliability data as we have been talking about, but it is an exceedingly difficult thing to explain the materials to the consumer in terms that he can understand—and I am not speaking merely of the elementary teacher and the high school teacher, although I am speaking of them, but I am speaking also of many counselors.

It happens that I have something to do with a counseling bureau at the university level. We have very intelligent people among our group, but one thing with which we have to struggle constantly is to get some common-sense interpretation of what a test means, plus what it shows in a case under consideration.

One other point: I think we are overdoing this matter of prediction as the objective of tests. I cannot agree with the implication that if you give an ACE, you will not need a test of English for the same population. We have plenty of evidence now, as far as that goes, that in some cases you will get about as good a correlation with the Q score as with the L score, but does that mean we should not have both? The answer depends on what you want to do with the scores. In many situations it is not merely a matter of looking ahead to see what the individual will probably do, but of helping him to get a program that will fit into his needs. Frequently we need tests that, so far as correlations are concerned, do not give us anything more than some test already administered has given, but do give us a better idea of what the individual is.

MR. LENNON: In all these castigations of the publisher and author, it seems to me I detect a premise that the more information which is provided about a test in the way of additional reliability and validity data,

norms, and so on, the more effectively the test can be used or will be used. While I think in principle one might agree with that, in practice it seems to me that that is only true within limits, and very decidedly is it true in relation to the competence and the sophistication of the test user.

We have to distinguish in our thinking among types of tests with respect to the amount and complexity of information which is likely to be used with wisdom, prudence, and good sense by the ordinary user of a given type of test.

To make this specific, we have had a little experience in providing several varieties of normative data for elementary school achievement tests, and a very common reaction is one of confusion on the part of the test user as to which kind of norms he should use or can use most effectively. It is the exception where we find that there is an enrichment of interpretation or an improvement in the use of the test by virtue of having the multiplicity and proliferation of normative data.

I would not want to suggest that we should discontinue any efforts to get additional data about tests; certainly not. But what I am saying is that to provide this additional information, which without doubt is very useful, maybe to people such as the people of this group, who are sensitive to the issues involved and can use this additional information well, the cost of the product must be increased, with no corresponding increase in the effectiveness with which the ordinary user can use the instrument—and I am thinking mostly of elementary school teachers—at an increased cost to them. There are limits to the amount of added material that the average user of given types of tests can digest and use well, and that is certainly one limiting factor, strictly from an economic standpoint, on the extent to which an author and publisher can go along with the desire for added information.

I want to add another observation on Dr. Conrad's suggestion that the publisher and the author accept a responsibility for preventing their tests from being abused, again thinking of the analogy with the drug dispenser or the pharmaceutical manufacturer who label their wares with cautionary notes, contraindications, and so on. Well, the number of ways in which a test or test score can be abused is legion. We certainly could make a catalog of the things that a person should *not* try to do with an elementary achievement test, an aptitude test, or an intelligence test. But I have a feeling that if we were to catalog all those things and say in the manual, "Now, don't try to do this, don't try to do the other with this test," we would be pretty much in the position of telling kids not to put beans in their ears. Most of them never would have thought of these things unless we had put it in their minds in the first place. There are certain gross mistakes which you know from experience are commonly made with certain types of tests. Well, naturally, you safeguard against those. But I

don't think you can go very far in trying to anticipate all the foolish or mischievous things that an uninformed test user is likely to do.

If I can add just one final word, again representing the vested interests, I would like to observe—and this stems from repeated experience in these sessions on what publishers ought to do—that I think the sessions would be far more productive if, instead of damning publishers and authors more or less generically, and then saying, "Oh, we don't mean you and we don't mean you," when Seashore and Durost get up—

CHAIRMAN BUROS: By the way, I didn't go down the line very far. I just warn you—

MR. LENNON: That's all right. Maybe you went down the line as far as we can go. I say, let's start with the next publisher or author on the list and say, "Now, this is what we mean," and if we are going to damn, let's damn in a concrete and not in a generic form, so that those responsible will have a chance to speak to the point.

CHAIRMAN BUROS: Dr. Conrad, do you care to comment on this at all?

DR. CONRAD: Well, I think that a distinction has to be made between the test user and the test purchaser. The elementary school teacher is frequently an agent who uses the test at the request of her principal or of her superintendent. It is quite possible that the teacher may be given a very much simplified cookbook to go with the test. But if the purchaser, the superintendent, or the principal, is going to make an intelligent choice, then he has to be educated to that intelligence. Unless the education is on a factual basis, tests will be sold by salesmanship instead of by merit. Therefore, I would suggest that while it is impossible, or not desirable, to put a whole book into each manual, nevertheless the information should be on tap as an assurance that the author knows what he is talking about and that the publisher knows what he is trying to sell.

CHAIRMAN BUROS: Do you want to talk to Dr. Conrad's remarks, Dr. Lorge?

DR. LORGE: I think the point which has just been made should be an indication of why we are having so much difficulty. If all the information that you needed for every test were provided in the manual, you would find that it would be impossible for any one human being to read all of it. I would suggest therefore that the next thing we do with every test manual is to include with it a reading test. We could find out the degree to which they can get the general, over-all view. What is this test about? They could get specific details of where the information about the test is in the manual, and then perhaps, check the kind of inferences the test user can make.

It seems, certainly, from the kind of talk that has been going on here today, that either the people do not know what they are talking about or they do not know how to move from one level to another. The idea of having an expectancy table for a

group such as this is a rather ingenious one, but I think if you knew what a correlation coefficient is, you should be able to move to an expectancy table, if you have any relationships reported. If the relationships are not of the quantitative but of the qualitative form, then I can see little point in the expectancy table.

It seems to me that we are trying to do too many things with too many people at the same time.

The last point I would like to bring to the attention of the group is that there is a serious problem here in public relations. We have assumed, because there is a group of people who call themselves ethical pharmacists, or ethical pharmaceutical manufacturers, that they are *ipso facto* ethical. Let me assure you that the number of people who misuse drugs is much greater than the number of people who misuse tests.

CHAIRMAN BUROS: I should like to check with the members of my panel to find out whether any of them have any further comments they would like to make. Would you like to make any comments, Dr. Dressel?

DR. DRESSEL: One final one, perhaps. I think this matter that came up about the use of an English test where a Scholastic Aptitude Test is already being used is a rather important thing. It fits right into the philosophy that I was trying to get out in my talk.

I cannot visualize the English Department on my campus being at all satisfied with any placement or sectioning based on a Scholastic Aptitude Test, even if it shows a higher correlation with grades in English than does an English test. I think we might also ask the question, If it becomes evident that a Scholastic Aptitude Test does correlate more highly with grades, may there not be something wrong with what is being done in that English course?

CHAIRMAN BUROS: Dr. Shaffer, do you care to make any additional remarks?

DR. SHAFFER: Throughout all of our discussion today there has been agreement with a pervasive but unacknowledged philosophy. We want tests that are practical, that will work, and that will correlate with definite criteria. The recent celebration of the ninetieth birthday of John Dewey brings to mind the source of our philosophy. We were all nurtured on John Dewey.

The two contrary attitudes about personality tests that I mentioned earlier arise from different basic sets of values. One view holds that personality can only be experienced esthetically, or perhaps even poetically; that there are basic truths about people that transcend criteria or practical applications. A similar philosophy can be found in some expressed views about achievement tests. It may be that English composition, or some other subject matter, is worth while in its own right, without regard to whether it correlates with or will predict anything else.

In both personality tests and achievement tests our support of a pragmatic position, our insistence that

tests must be useful and valid in the sense that they correlate with criteria, is an expression of our philosophic values. We like those values, and acknowledge our intellectual descent from John Dewey and William James. But other cultures and other eras may adopt other philosophies, and we must admit that we do not necessarily have the ultimate truths.

CHAIRMAN BUROS: I wonder if there is a chance, Dr. Kelley, that you would care to make a few remarks?

DR. KELLEY: Naturally, I am very much interested in almost everything that has been said. Practically all of it has touched me personally in one way or another. I certainly added my share of matches in trying to put the heat on, and Walter Bingham has been one of my supporters in the matter. We have tried to put the heat on, and tried to get not only authors and publishers but users as well to depend upon information about reliability and validity that they were not accustomed to depend upon.

I don't know how successful we have been. Certainly, in a group like this, I know without a doubt that you people have a dependence upon phenomena of reliability and validity that could not be paralleled when the testing movement started, not, as Oscar Buros said now, because I think it started some little time ago. It started back in the days of Thorndike and Hillegas and even earlier. But you depend upon evidence of reliability and validity, probability distributions, and so forth, in a degree and to an extent which, to me, is very encouraging. It is a degree and an extent that did not exist not so very many years ago. If you depend upon it, it is going to spread. It is going to spread to publishers and test users.

Now, just how much the user can digest, I do not know. I would like to mention one little device, and then ask Dr. Durost, perhaps, whether the public has digested it. In an early form of the Stanford Achievement Test, we had a profile, and we put on that profile a little bar that indicated the size of the probable error of a score. That was the first time that was used, and I remember that the publisher, which was the World Book Company, said, "Well, we don't want that. Nobody knows what it means." I said, "It is rather inconspicuous; it isn't going to be very serious, so just put it in." We had these probable-error bars on the profile, and the score at this point had this probable error, and the score at that point had this one. Today, I would like to do it over again and have standards errors, but, anyway, we had those. I would like to ask Walter Durost, has that been digested?

DR. DUROST: I am afraid that the answer to that is definitely, "No," but the bar is still there. We hope that some day somebody will pay attention.

DR. KELLEY: I think, perhaps, Dr. Buros, I will not make any further remarks. One can talk almost endlessly, but we have had such a fine session of discussion that it is not necessary to go on at all. I thank you.

CHAIRMAN BUROS: Thank you very much, Dr. Kelley. I should like to leave just one thought with you—and that is, there is a danger here, where we set up the objective of what we should like test publishers and test authors to supply. We are mentioning all of these particular things which we would like—well, no, we are not mentioning all, but quite a large number. Just because you cannot supply us with all of this information is certainly no argument that you should continue your past practice. You can make progress in the direction of giving us more adequate information than you are doing now, and I should be glad to make suggestions to publishers as to how they can give us some information, which they possess in their offices, at very little, if any, extra expense.

It has been very difficult for me to keep out of this afternoon meeting more than I have. I just seem to want to talk all the time. I should be very glad to have one of these bureaus of test evaluation set up so that it would permit me to take a rest and not try to put out any more *Mental Measurement Yearbooks*. I should be very glad to go out of business.

Appendix

PARTICIPANTS—1949 INVITATIONAL CONFERENCE ON TESTING PROBLEMS

ADAMS, Elizabeth C., Educational Testing Service
ADAMS, Joe Kennedy, Bryn Mawr College
ADKINS, Dorothy C., University of North Carolina
ALLART, Shirley J., Educational Testing Service
ANASTASI, Anne, Fordham University
ANDERSON, Rose G., Psychological Service Center
ANDERSON, Roy, North Carolina State College
ANDREWS, T. G., University of Maryland
ANGOFF, William, Educational Testing Service
ARMSTRONG, Hubert C., Harvard University
ARNOLD, Samuel T., Brown University
ARSENIAN, Seth, Springfield College

BEARDSLEY, Seymour W., Polytechnic Institute of Brooklyn
BECK, Hubert P., City College of New York
BELLOWS, Roger M., Wayne University
BENNETT, George K., Psychological Corporation
BENSON, Arthur L., Educational Testing Service
BERGESEN, B. E., Jr., Educational Testing Service
BERNSTEIN, Alfred J., Queens College
BINGHAM, Walter V., Washington, D.C.
BLOOM, Benjamin S., University of Chicago
BOOKSTAVER, J. B., Teaneck, New Jersey
BOWLES, Frank H., College Entrance Examination Board
BRANDT, Hyman, Personnel Research Section (AGO)
BRANSFORD, Thomas L., New York State Civil Service Department
BRIDGES, Claude F., World Book Company
BRISTOW, William H., New York City Board of Education
BRYAN, Miriam M., Silver Burdett Company
BUCKTON, LaVerne, Brooklyn College
BURNHAM, Paul S., Yale University
BUROS, Oscar K., Rutgers University

CAROL, Bernard, Metropolitan Life Insurance Company
CARROLL, John B., Harvard Graduate School of Education
CARSTATER, Eugene D., Navy Department, Washington
CASE, Ethel E., Polytechnic Institute of Brooklyn
CAYNE, Bernard, Educational Testing Service
CHAUNCEY, Henry, Educational Testing Service
CHURCHILL, Ruth D., Antioch College
COHEN, Elizabeth, Educational Testing Service
COHN, Fannie M., International Ladies Garment Workers' Union
CONRAD, Herbert S., U.S. Office of Education, Washington
COPELAND, Herman A., Atlantic Refining Company
CORNELL, Ethel L., New York State Education Department
COWARD, Ann F., Educational Testing Service
COWLES, John T., Educational Testing Service
CRANE, Percy F., University of Maine
CRISSY, William J. E., Queens College
CROSS, Mary, Educational Testing Service
CUTTS, Norma E., Yale University
CYNAMON, Manuel, Brooklyn College

DAVIS, Allison, University of Chicago
DAVISON, Hugh M., Pennsylvania State College
DERRICK, Clarence, Educational Testing Service
DETCHEN, Lily, Pennsylvania College for Women
DIEDERICH, Paul, Educational Testing Service
DION, Robert, California Test Bureau, New Hampshire
DOPPELT, Jerome E., Psychological Corporation
DRAGOSITZ, Anna M., Educational Testing Service

83

DRESSEL, Paul L., American Council on Education
DUBOIS, Philip H., Washington University
DUROST, Walter N., Boston University
DUVALL, Ellen N., Richmond, Virginia
DYER, Henry S., Harvard University

EBEL, Robert L., State University of Iowa
EDGERTON, Harold A., Richardson, Bellows, Henry & Co.
EYSENCK, H. J., University of London, England

FAN, Chung-Teh, Educational Testing Service
FEIFEL, Herman, Personnel Research Section (AGO)
FELS, William, College Entrance Examination Board
FINDLEY, Warren G., Educational Testing Service
FINKLE, Robert B., Metropolitan Life Insurance Company
FISHER, Kathryn B., Rinehart and Company
FLANAGAN, John C., University of Pittsburgh
FLEMMING, Cecile W., Burton Bigelow Organization
FLEMMING, E. G., Burton Bigelow Organization
FOLEY, John P., Psychological Corporation
FREDERIKSEN, Norman O., Educational Testing Service
FRENCH, Benjamin J., New York State Civil Service
FRENCH, John W., Educational Testing Service
FUCHS, Edmund F., Personnel Research Section (AGO)

GALLAGHER, Henrietta, Educational Testing Service
GARDNER, Eric F., Syracuse University
GERBERICH, J. Raymond, University of Connecticut
GILMORE, John V., Boston University
GORDON, Hans C., Pennsylvania Board of Education
GREEN, Bert F., Jr., Educational Testing Service
GREENE, Edward B., Wayne University
GULLIKSEN, Harold, Educational Testing Service

HAGEN, Elizabeth P., Hunter College
HAGGARD, Ernest, University of Chicago
HAGGERTY, Helen R., Personnel Research Section (AGO)
HAGMAN, Elmer R., Connecticut Board of Education
HASTINGS, J. Thomas, University of Illinois
HAUSRATH, Alfred H., Jr., Educational Testing Service
HAVIGHURST, Robert J., University of Chicago
HEATON, Kenneth L., Boston University
HESS, Robert, University of Chicago
HESTON, Joseph C., DePauw University
HIERONYMUS, Albert N., State University of Iowa
HOFFMAN, E. Lee, Educational Testing Service
HOGAN, Ralph M., Navy Department, Washington
HORST, Paul, Educational Testing Service
HORTON, Clark Willis, Dartmouth College
HOSMAN, Captain Richard S., U.S.A.F. Special Staff School, Alabama
HUDDLESTON, Edith M., Educational Testing Service
HURD, Cuthbert C., International Business Machines Corp.

ILLSON, Murray, New York Times

JACOBS, Robert, Educational Records Bureau
JASPEN, Nathan, Pennsylvania State College
JOHNSON, A. Pemberton, Educational Testing Service
JONES, Galen, U.S. Office of Education, Washington

KABACK, Goldie R., City College of New York
KARLIN, John E., Bell Telephone Laboratories
KATZEL, Raymond A., Syracuse University
KELLEY, Truman L., Harvard University
KINNEY, Lucien B., Stanford University
KLINE, William E., Educational Testing Service
KOGAN, Leonard S., Community Service Society
KREILING, Dorothy, Eagle Avenue School, New York
KURTZ, Albert K., Pennsylvania State College
KUSHNER, Rose E., Teachers College, Columbia University

LANGMUIR, Charles R., Syracuse University
LANNHOLM, Gerald V., Educational Testing Service
LAPHAN, Robert, Burroughs Adding Machine Company
LAYTON, Wilbur L., University of Minnesota
LEACH, Kent, University of Michigan
LENNON, Roger T., World Book Company
LEVERETT, Hollis M., American Optical Company
LINDQUIST, E. F., State University of Iowa
LITTERICK, W. S., Peddie School
LONG, Louis, City College of New York
LORD, Frederic M., Educational Testing Service
LORGE, Irving, Teachers College, Columbia University
LURIE, Walter A., National Community Relations Advisory Council

MCCALL, William C., University of South Carolina
MCNAMARA, Walter J., International Business Machines Corp.
MCNEMAR, Quinn, Stanford University
MCQUITTY, John V., University of Florida

MACPHAIL, Andrew H., Brown University
MANUEL, Herschel T., University of Texas
MARTIN, Lycia O., State Teachers College
MEYER, Lorenz A., Veterans Administration, Washington
MICHAEL, William Burton, Princeton University
MILLER, Ruth, Educational Testing Service
MOLLENKOPF, William G., Educational Testing Service
MOSIER, Charles I., Personnel Research Section (AGO)
MUIRHEAD, Peter P., New York State Education Department
MYERS, Charles, Educational Testing Service

NOGEE, Philip, Educational Testing Service
NOLL, Victor H., Michigan State College
NORTH, Robert D., University of Kentucky

OLSEN, Marjorie, Educational Testing Service
ORLEANS, Beatrice S., Bolling Air Force Base
ORLEANS, Jacob S., City College of New York

PACE, C. Robert, Syracuse University
PASHALIAN, Siroon, New York University
PEASE, Katharine, Barnard College
PERRY, William D., University of North Carolina
PETERSON, Donald A., Life Insurance Agency Management Association
PLUMLEE, Lynnette B., Educational Testing Service
POTTS, Edith M., Psychological Corporation
PRESTON, B. Braxton, Educational Testing Service

REMMERS, Hermann H., Purdue University
RETHOLTZ, George, Bolling Air Force Base
RICCIUTI, Henry, Educational Testing Service
RICKS, James H., Jr., Psychological Corporation
RIEGEL, Eleanor, Educational Testing Service
RIMALOVER, Jack K., Educational Testing Service
ROCK, Robert T., Fordham University
ROESSLE, Robert L. B., Standard Oil Company of New Jersey
RULON, Phillip J., Harvard University

SCATES, Douglas E., American Council on Education
SCHRADER, William B., Educational Testing Service
SCHROEDEL, Edward C., International Business Machines Corporation
SCHULTZ, Douglas G., Educational Testing Service
SEASHORE, Harold G., Psychological Corporation
SHAFFER, Laurance F., Teachers College, Columbia University
SHARP, Catherine G., Educational Testing Service
SHAYCROFT, Marion F., National League of Nursing Education
SIEVERS, Frank, U.S. Office of Education, Washington
SMITH, Allan B., University of Connecticut
SMITH, Denzel D., University of Maryland
SMITH, M. Brewster, Vassar College
SPANEY, Emma, Queens College
SPEER, George, Illinois Institute of Technology

SPENCE, Donald P., Educational Testing Service
SPENCE, Ralph B., Teachers College, Columbia University
SPENCER, Lyle M., Science Research Associates
STALNAKER, John M., Educational Testing Service and Illinois Institute of Technology
STEPHENSON, William, University of Chicago
STEWART, Naomi S., American Jewish Committee
STRONG, E. K., Jr., Stanford University
SULLIVAN, Richard H., Educational Testing Service
SUPER, Donald E., Teachers College, Columbia University
SWINEFORD, Frances, Educational Testing Service
SYMONDS, Percival M., Teachers College, Columbia University

THORNDIKE, Robert L., Teachers College, Columbia University
TIEDEMAN, David V., Harvard University
TORGERSON, Warren S., Educational Testing Service
TRAVERS, Robert M. W., Hunter College
TRAXLER, Arthur E., Educational Records Bureau
TRIGGS, Frances O., Committee on Diagnostic Reading Tests, Inc.
TROYER, Maurice E., Syracuse University
TUCKER, Ledyard R, Educational Testing Service
TURNBULL, William W., Educational Testing Service

WALKER, Helen M., Teachers College, Columbia University
WATSON, Fletcher G., Harvard Graduate School of Education
WATSON, Walter S., Cooper Union
WATT, R. R. G., University of Southern California
WECHSLER, David, Bellevue Psychiatric Hospital
WEITZ, Henry, University of Delaware
WENZEL, Bernice M., Barnard College
WESMAN, Alexander G., Psychological Corporation
WHITNEY, Alfred G., Life Insurance Management Association
WILSON, John T., Navy Department, Washington
WINGO, Alfred L., Virginia State Department of Education
WOLMAN, Benjamin, Tel-Aviv, Israel
WOOD, Ray G., Ohio State Department of Education
WRIGHTSTONE, J. Wayne, New York Board of Education

ZUBIN, Joseph, Columbia University